Exploring log-normal distributions in nascent entrepreneurship outcomes:

Exploring log-normal distributions in nascent entrepreneurship outcomes:

Rodriguez Hernandez, Ivan Francisco

Award date:
2019

Awarding institution:
University of Bath

Exploring log-normal distributions in nascent entrepreneurship outcomes: International comparisons and agent-based modelling.

Iván F. Rodríguez Hernández

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

School of Management

June 2019

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

Candidate signature: 

Declaration of authorship

I am the author of this thesis, and the work described therein was carried out by myself personally.

Candidate signature: 

To my grandfather, José Hernández de León.

TABLE OF CONTENTS

TABLE OF FIGURES

ABSTRACT

Recently, Joo, Aguinis and Bradley (2017), using a novel distribution pitting technique, have found that the exponential tail distributions-- exponential and power law with an exponential cut-off -- and their generative mechanism – namely, incremental differentiation -, are the most frequent distribution in many individual outputs across different organizations, sectors, jobs and activities.

However, this may not be totally accurate in nascent entrepreneurship processes: the first section of this research shows that the **lognormal distribution** in entrepreneurial outcomes seems predominant throughout the different panels – i.e., longitudinal studies - in different countries. We have studied those in which the datasets are in the public domain: Australia, Sweden, US PSED I & II (Reynolds, 2017b). The power law distribution with an exponential cut-off may also be a plausible fit in some particular panel outcomes variables. A definitive conclusion regarding which of these two distributions may be the better fit will require the analysis of the rest of 14 still ongoing longitudinal projects around the world. The pervasiveness of lognormality offers relevant clues to understand nascent entrepreneurial processes, their generative mechanism, and it will offer strategies to allocate resources to foster and promote new entrepreneurial ventures.

The second section of this research is the design and implementation of **a baseline agent-based model** as a research tool, "*A nascent entrepreneurial agent-based model*". Inspired by previous simpler entrepreneurial models, our model introduces new layers of complexity, making possible parametrization and calibration. This baseline model, initially with parameters similar to the public available panel datasets -- Australia, Sweden, US PSED --, is able to generate the patterns that were found in the empirical results: the heavy-tailed distributions.

Although PSED-type of longitudinal panels have been performed in more than a dozen countries, their results and datasets are not publicly available yet. This base model is, therefore, flexible in order to be easily adapted to each of the empirical dataset under study. The model, at this initial stage, has not been fully parametrized and calibrated for any specific country. The baseline model takes the main parameters from the datasets available as examples, in order to show that multiplicative processes --as main generative mechanism-- are able to simulate the empirical patterns.

The baseline model is designed as a **research tool** to experiment and to help entrepreneurship researchers to test their theories, and for exploring in more detail the mechanisms involved in the emergence of new ventures. The baseline model and its background documentation will be openly available to the research community in two major agent-based repositories. Taking this baseline model as a “backbone”, researchers can change parameters, agents, behaviours, schedules or global variables for their own theory building or calibration of their specific country’s simulation.

Keywords

Heavy-tailed distributions

Power-law distributions

Generative processes

Fitting procedures

Nascent venturing processes

Agent-based modelling

1. INTRODUCTION: THE PARADIGM SHIFT

1.1 THE CONCEPTUAL SHIFT FROM GAUSSIAN DISTRIBUTIONS TO HEAVY-TAILED DISTRIBUTIONS IN ORGANIZATIONAL RESEARCH: ANTECEDENTS.

Although the concept of “paradigm” is highly controversial (Tasaka, 1999), the academic community would agree that complexity – or “complexity sciences” or “complexity theory” - can be considered an emerging post-Newtonian paradigm (Kuhn, 1996). It tries, from a somehow unifying point of view, to address specific phenomena that occur in systems constituted by many subunits, drawing on methods, concepts and tools from nonlinear dynamics, statistical physics, probability and information theory, data analysis, networks and numerical and agent-based simulations (Nicolis and Rouvas-Nicolis, 2007). “Complexity” studies complex systems. It is an interdisciplinary domain that tries to explain how large numbers of simple entities organize themselves, without any central controller, creating patterns, using information, and, in some cases, able to learn and evolve. Complex systems may include ant colonies, immune systems, brains, markets and economies (Mitchell, 2009).

Currently, complexity - its methods, concepts and tools - is ubiquitous in natural and social sciences, although the development of its theory is in an incipient stage and a general unified framework across disciplines is still missing (Sporns, 2007). Newman (2011) would argue that there is not a “general theory of complex systems”, but rather a body of knowledge with different “theories” not fully integrated yet (Newman, 2011). The mathematician Steven Strogatz has suggested that science has not yet developed the right concepts and mathematical tools to address complex systems and to formulate and describe the different forms of complexity that are seen in nature and societies; that something such as a conceptual equivalent to calculus is missing, an “ultracalculus” – as he calls it - able to

model the multiple interactions of a complex system (Strogatz, 2004). In any case, there are properties common to all complex systems, and several universal complex systems principles have been proposed. Some examples are: the universal properties of chaotic systems, the principles of self-reproduction (John von Newman), the principle of balancing exploitation and exploration (John Holland), general conditions for the evolution of cooperation (Robert Axelrod), the principle of computational equivalence (Stephen Wolfram), the principle of preferential attachment as a general mechanism for the development of real-world networks (Albert-Laszlo Barabasi and Reka Alberts), etc. (Mitchell, 2009).

In the last 20 years, Complexity Science, originally developed mainly in the context of the mathematical description of natural dynamical systems, has been progressively used in Organization Studies, Economics and Management. In related fields, such as economics or finance, complexity science has an increasing and stronger presence than in Management or Organizational Studies (especially in financial time series analysis) (Mantegna and Stanley, 1999; Sornette, 2004; Easley and Kleinberg, 2010). In economics, for example, the works of Harvard Economist Ricardo Hausmann and MIT's physicist Cesar A. Hidalgo, at the Observatory of Economic Complexity (Harvard-MIT), have introduced the concepts of Economic Complexity and Economic Complexity Index (Hausmann, Hidalgo et al., 2011). As it has occurred in the past, gradually, cross-fertilization among disciplines and the borrowing of theory and analytical techniques from one to another is becoming common practice (for example, how XIX century economics mimicked XIX physics concepts) (Whetten, Felin & King, 2009).

Now, it is becoming quite frequent to find Complexity Science-based papers on *Organization Science* (the first monographic issue on this topic was edited in 1999), *Academy of Management* publications, and many other relevant journals in the field. *The SAGE Handbook of Complexity and*

Management tried to recapitulate the work published on this specific approach during the last years, and to draw the possible future lines of research (Allen et al., 2011).

Initially, in the 1990s, the Complexity approach in management research tried to explicate the new concepts and terminology (chaos, fractals, emergence, nonlinear dynamics, networks, self-organization, complex adaptive systems, etc.), to describe the methods, and to introduce the "New Science" (Wolfram, 2002) to the Management academia and practitioners, explaining its potential implications (Maguire, 2006). Most of the works during that decade was descriptive, it was scarce in empirical studies, and very few works developed theory or models (Maguire, Allen and McKelvey 2011). McKelvey (1999) pointed out that without sound complexity applications rooted in empirical data and solid theoretical foundations there were a risk of turning Complexity Science applied to management into an "another management consulting fad". A review conducted by Maguire and McKelvey (1999) in 1999 found that the books published until then on Complexity and Management were mostly and merely "metaphorical" and thought-provoking, but without presenting the full toolkit of complexity methods (Maguire and McKelvey, 1999).

However, soon, the mathematical tools and the Complexity Science methods also demonstrated their capabilities in Management and Organizational research, mainly in the introduction of a new mathematical formalism and in the development of its new computational modelling techniques: cellular automata, genetic algorithms, neural networks, Kauffman's NKCS "fitness landscape", Agent-Based Modelling, etc. In the special issue of the journal *Organization Science* devoted to the application of Complexity Theory to this field (*Organization Science* Vol. 10, No 3 May-Jun 1999), Anderson (1999) published - under the section "Perspective" - a call for the need of the organizational scholars to understand at a high level how to use these new computer models given their potential to open new

perspectives on organizational life (Anderson, 1999). This Special Issue of 1999 was one of the major milestones in the introduction of Complexity Science to the Management scholars. In the same year, 1999, the first journal on complexity science and organization studies appeared, *Emergence* - now called *Emergence: Complexity and Organization: E:CO*) - aiming to build empirical and theoretical solid foundations in this incipient and interdisciplinary field.

By 2006, Maguire et al. (Maguire, 2006) found and reviewed around 331 references in Organization Studies and Management research using complexity concepts and methods, which, compared with the enormous amount of references based in Complexity Theory in other disciplines – mainly in natural sciences -, shows that this approach was still immature in the fields of Management and Organizational Sciences.

In 2009, ten years later after the special issue on Complexity and Organizational research, the journal *Organization Science* revisited again the interdisciplinary area of Complexity Science and Organization Studies. The paper in the section “Perspective” – which points out the more promising future lines of research - was authored by Andriani and McKelvey (2009). Andriani and McKelvey (2009) proposed to redirect Organization Science research toward Pareto distributions. They thought that, although there are some topics in which normal/Gaussian/bell shape distributions fit properly and have an appropriate application, the discovery of the Pareto rank/frequency distribution – or other types of heavy tail distributions - in organizational datasets make necessary to incorporate its specific statistics.

Many of the Gaussian statistics are not meaningful addressing these heavy-tailed distributions in social science datasets. Pareto means are unstable or non-existent. The Paretian distribution has long and “fat/heavy tails”. These power laws show potentially infinite variance: the variance may

cross many orders of magnitude (from the revenues of a small store in a countryside village – in the range of thousands of dollars - to the billions of dollars revenues of Wall-Mart globally – ranges can go across 11 magnitudes -). The confidence intervals are thus less significant. Gaussian statistics also miss "key extreme outliers", where significant events occur, such as the emergence of "Facebook" or "Twitter".

What do Pareto/heavy-tailed distributions say from a theoretical and causal point of view? What is the difference with the Gaussian, normal distribution? According to these authors, "*the difference lies in assumptions about the correlations among events*" (Andriani and McKelvey 2009, p. 1055). From a Gaussian perspective, data are independent-**additive**, and generate the normal Gaussian distribution. On the order hand, when causal elements are independent-**multiplicative**, they show a lognormal distribution (another kind of heavy-tailed distribution). However, when events are interdependent, interactive, or both, Pareto distributions emerge.

Andriani and McKelvey (2009) complained that most of the statistics in organization science is based on the Gaussian, normal distribution scheme and denounced that many social researches had decided to ignore other distributions such as the Paretian ones. They claimed that Pareto distribution and its specific statistics are commonly unknown to most quantitative organizational researches. On the other hand, the use of certain mathematical tools is not neutral. Any mathematical tool has its own philosophical and methodological background. Gaussian statistics is related to "linear science and linear way of thinking" – as opposed, for example, to the Poincaré's chaos mathematics -.

> *"The adoption of normal distribution statistics carries a heavy burden of assumptions. Reliance on linearity, randomness, and equilibrium influences how theories are built, how legitimacy is conferred, and how research questions are formulated" (Andriani and McKelvey 2009, p. 1053).*

This consideration is very important for the authors because "*ignoring power-law effects risk drawing false conclusions and promulgating useless advice to practitioners*" (Andriani and McKelvey 2009, p. 1053). Real organizations and real managers live in a world with interdependent events (not of the Gaussian hypothetical independent ones).

But how to explain the presence of heavy tailed distributions in organizational data? How to explain the scalability (fractal geometry), the self-similarity (McKelvey, Lichtenstein and Andriani, 2012)? What are the forces that cause the scalability patterns or scaling laws? Andriani and McKelvey (2009) suggested 15 Scale Free theories that can be applied to organizations and that would explain the power-law scaling behaviour of those systems. We will explore some these scale-free theories in subsequent sections of this document. For Andriani and McKelvey (2009), "the power law signature" is the best evidence of emergence, which operates in different organizational dimensions.

1.2. THE DISCOVERY OF NON-NORMAL AND HEAVY RIGHT-TAILED DISTRIBUTIONS IN THE FIELD OF ENTREPRENEURSHIP

B.B. Lichtenstein has extensively reviewed the contribution of the Complexity Theory to understand the emergence of firms and he has pointed out the areas that do still need further research (Lichtenstein, 2011). Bill McKelvey has even postulated the possibility of developing an entrepreneurship theory using the Complexity Science corpus and its specific methodological and mathematical tools (McKelvey, 2004). Although the research using the tools of Complexity Science is still embryonic in entrepreneurship studies, it is becoming more common. The Best-Paper of

2013 *Academy of Management Annual Conference Proceeding* was precisely based on the application of complexity science theory and tools to the study of the emergence of new ventures (Crawford and Lichtenstein, 2013).

One of the conceptual keys of entrepreneurship, at all different levels of analysis, is the concept of **emergence** (Lichtenstein, 2011). This emergence can be of firms, technologies, networks, clusters, new markets, industries, institutions, etc. Complexity Science offers useful models to understand the emergence of new patterns and structures in the natural and social world. Hence, some of the tools used by Complexity Science may be indispensable to study entrepreneurial emergence.

Lichtenstein (2011) has made quite an extensive compilation of the different complexity science approaches applied in entrepreneurship research to explain the phenomenon of emergence and the entrepreneurial processes (Lichtenstein, 2011). However, as in the field of Management, most of the research on complexity and entrepreneurship has been "metaphorical". Lichtenstein (2011) distinguished four types of contributions to entrepreneurship and entrepreneurial emergence, where Type I consists in using complexity as "metaphor", Type II is defined as "discovering" complexity, Type III is modelling complexity - it includes the studies using agent-based modelling and simulations -, and Type IV is related to "generative" complexity. Lichtenstein (2011) studied and classified the 28 published papers that specifically apply Complexity Science to entrepreneurship. It is interesting to notice that in Lichtenstein's review that spans for almost 20 years, there is a small number of papers and research works conducted with this interdisciplinary approach – just 28 papers, few of them empirical -, showing somehow "a gap in the field", an underdeveloped line of research. Given that **emergence** is a core theme in entrepreneurship, and the theoretical and methodological power of the Complexity Science tools dealing with emergence that so fruitfully has been

applied in other disciplines – from astrophysics to neurobiology -, this lack of entrepreneurial research using Complexity Science is perplexing (Lichtenstein, 2011). It is also strange the lack of entrepreneurial research on **emergent networks**, given that networks are so relevant for entrepreneurial development and the success of network science in other fields (Lichtenstein, 2011).

The emergence of new ventures is one of the current central themes in entrepreneurship research (Gartner, 1985; Westhead and Wright, 2013). However, many scholars have pointed out the lack of enough knowledge about its causes, effects, and processes, and the lack of a global perspective on this complex and multi-dimensional phenomenon (Leitch et al., 2010). An adequate knowledge of the phenomenon of emergence of new firms is necessary given the importance of entrepreneurship at different economic and social levels of an economy (Amorós and Bosma, 2014). Furthermore, entrepreneurship activity — the process of starting and establishing a new business — has already demonstrated to be essential for the development, growth and prosperity of nations, increasing the competitiveness of an economy, creating jobs, reducing unemployment, developing innovation, and fostering economic and social mobility (OECD, 2007; Naudé, 2010; Baumol and Schilling, 2008), in particular by that small proportion of high-performing new ventures – the high impact firms or “gazelles” - that are the driver of the majority of innovation, wealth creation, and new job generation (Nightingale and Coad, 2014).

However, there is yet neither a comprehensive theory of creation of new ventures nor a consolidated praxis that help the entrepreneurial process (Headd, 2003; Crawford and McKelvey, 2012; Westhead and Wright, 2013; Crawford et al., 2014; Crawford et al., 2015). For example, why do very few new firms survive after three years? What is the underlined dynamics that produces such results and these high new firm closure rates (Westhead and Wright, 2013)? Is the emergence of new firms a “random

walk" process, a "game of chance", a case of "Gibrat's law", "a variant of Gambler's Ruin in which performance is random but where survival merely depends on access to resources" as Coad and colleagues propose (Coad et al., 2013)? Would a comprehensive theory of the entrepreneurial processes help us to 1) explain it, 2) foster it, and 3) mitigate the economic, social and emotional damages of firm closure?

On the other hand, an accurate knowledge of entrepreneurial processes may enhance the effect of public policies for the promotion of creation of new ventures avoiding wasting scarce public resources, and making those policy interventions effective and efficient (Pons Rotger, Gørtz and Storey, 2012). Currently the cost-benefit analysis of these policy interventions has been proven to be extremely difficult to ascertain both in its overall effectiveness, and in the effectiveness of its diverse elements (Lundström et al., 2014). Should policy support focus on promoting a large number of new firms, or concentrate the resources in fewer companies but with more wealth creation potential? How should this potential be measured? The understanding of the mechanisms of firms' emergence may help to implement a better cost-benefit analysis and effectiveness evaluation, as well as a better selection of the better new firms to invest in (Westhead and Wright, 2013; Arshed, Carter and Mason, 2014). **The objective of this PhD research is, precisely, to explore these plausible generative mechanisms that may explain the emergence of new ventures and the idiosyncratic features of their outcome datasets -- the presence of heavy-tailed distribution --.**

After the analysis of three panel (- longitudinal -) studies referred to the creation and emergence of new ventures in the United States of America, Crawford and McKelvey found that nascent entrepreneurial outcome variables such as numbers of employees or revenues, follow long-tail distributions that they identified – with the fitting software techniques available then - as power-law distributions (Crawford and McKelvey, 2012;

Crawford et al., 2014; Crawford et al., 2015). Initially, Crawford, McKelvey and Lichtenstein (2014) were able to pinpoint the long-tail distributions in outcome variables such as nascent ventures' numbers of employees and annual revenues in several datasets. Subsequently, Crawford, Aguinis, Lichtenstein, Davidsson, and McKelvey (2015) detected these long-tail distributions not only in outcome variables (number of employees, revenues, etc.) but also in input variables such as entrepreneurial resources, entrepreneurial activities, etc.

Although, by then, there were some studies on the application of Complexity Sciences methods and tools in Management, Organizational and Business Studies (Maguire et al., 2006; Allen et al., 2011), and an important tradition on the study of highly skewed distributions in economics and finance, **Crawford and McKelvey's (2012) paper was the first study to specifically address the presence of power-law, i.e. heavy tailed, distributions in nascent entrepreneurship datasets.** That is, only a very small number of nascent entrepreneurs become better-off over time, while most of them have less entrepreneurial success, attaining smaller outcomes. Later on, Crawford et al. extended and deepened this line of research in Crawford et al. (2014) and Crawford et al. (2015), where they introduced a new theoretical approach and proposed several alternative methodological techniques more suitable for addressing heavy tailed distributions (Bayesian statistics, agent-based computational modelling, etc.).

Several studies had already shown that the size (numbers of employees) distributions of already established firms were well described by a power law (a Zipf's law). These power law distributions also hold for other different measures, such as assets or market capitalization in the United States (Axtell, 2001; Fujiwara, 2004; Gabaix, 2009) and, even for those measures, for distributions in other countries (Gabaix, 2008). The extinction of firms also seems to follow a scaling invariant distribution,

another power law (Cook and Omerod, 2003; Di Guilmi, Gallegati, and Ormerod, 2004).

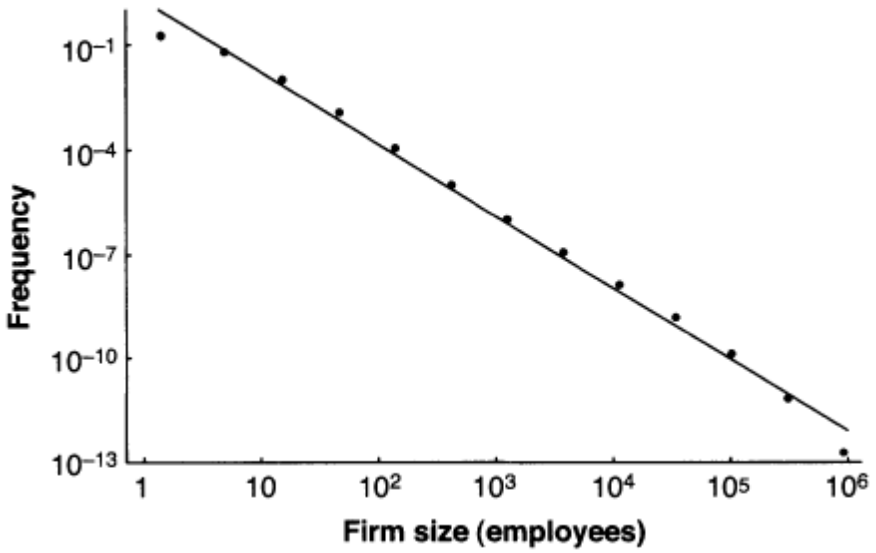


FIGURE 1 - ***FIGURE FROM AXTELL (2001, P. 1819), REPRESENTING THE FREQUENCY OF U.S. FIRMS SIZE BY EMPLOYEES PLOTTED IN LOG-LOG AXES, CORRESPONDING TO A POWER LAW DISTRIBUTION WITH EXPONENT 1.059 (APPROXIMATELY A ZIPF'S LAW). DATA ARE FOR 1997 FROM THE UNITED STATES CENSUS BUREAU.***

According to Crawford and McKelvey's initial empirical analysis (2012), power laws were ubiquitous in the six outcomes tested within three American entrepreneurial longitudinal datasets. They proposed that these results may offer an empirically validated comprehensive theory of the emergence of new firms and ventures. Indeed, literature on entrepreneurial creation continuously mentions the difficulties of developing a comprehensive theory of new ventures' dynamics (Leitch, Hill and Neergaard, 2010; McKelvey and Wiklund, 2010).

Crawford and McKelvey's contribution, with their heavy-tailed distribution analyses, opened a new line of research that may explain the skewed outcomes observed in data related to the emergence of new firms. In that paper of year 2012, and in the subsequent of years 2014 and 2015, Crawford et al. studied entrepreneurial outcomes embracing the line of

research pointed by Anderson (1999) and Andriani and McKelvey (2009) in their seminal papers published in *Organization Science*, in which they proposed to focus on Pareto's statistics, based on the study of interdependence and interconnection, rather than on the traditional and standard Gaussian approaches, linked to events completely independent and identically distributed. The Gaussian (normal) perspective and methods may not be able to explain the highly skewed distributions in emerging firms which have to deal with such a diverse reality, from thousands of retail high street stores to an extreme event, such as the uniqueness of the emergence of Amazon or Google (McKelvey, Lichtenstein and Andriani, 2012).

As we mentioned above, Andriani and McKelvey (2009) showed that heavy tails are present in the organizational world, pointing out to Pareto rank/frequency distributions, fractals, scale-free phenomena, and nonlinear organizational dynamics. They reviewed the presence of power laws in social networks, industry sectors, growth rates of firms, bankruptcies, transition economies, profits, sales decay, economic fluctuations, intra-firm decisions, consumer sales, salaries, size of firms, ecosystems, sector networks, etc., introducing more than a hundred of different kinds of power laws in organizational setting, suggesting that these Pareto rank/frequency distributions are more common that it seems to be, and much more relevant for organizational research and practice than they are considered now.

When plotted in double-log scales, a Pareto rank/frequency distribution appears with an inverse sloping straight line (what it is called the "inverse power-law signature").

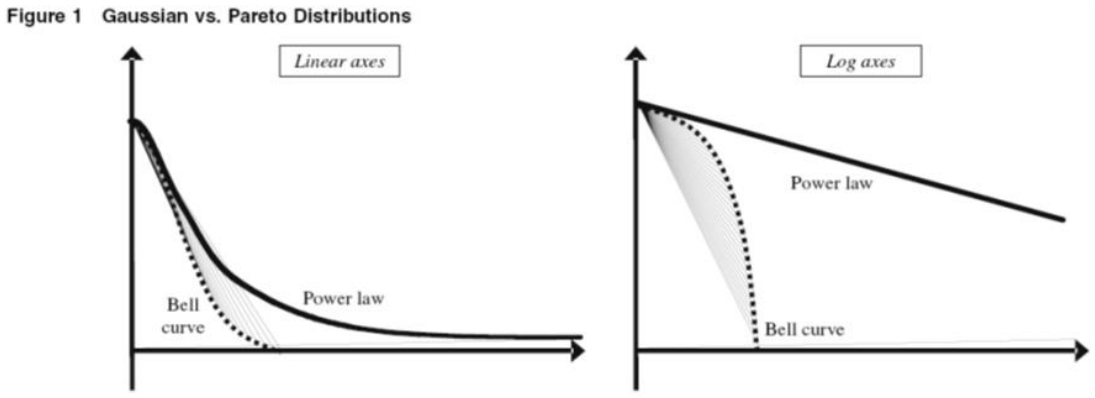


FIGURE 2 - *FIGURE FROM ANDRIANI AND MCKELVEY (2009, P. 1056).*

The overwhelming presence of heavy tails distributions, also discovered in entrepreneurship, challenges the traditional "normal distribution", the "Gaussian bell curve" researchers' mind-set and may force to change research methods and theoretical assumptions in the field (Crawford et al., 2015). Crawford and McKelvey (2012) claimed that their findings would be the foundation of the building of a new theory of entrepreneurial emergence using complexity science: they "*provide an empirically validated basis for a comprehensive theory of venture growth*" (Crawford and McKelvey, 2012, p. 2). Crawford et al. (2015) enumerated the most relevant potential generative causal processes that may produce heavy tailed distributions. However, they did not intent to enter in further theoretical development that would have been necessary to explain the presence of heavy tailed distributions and to describe the processes that produce their emergence. What could these processes be in entrepreneurship?

1.3. THE ANTECEDENTS OF OUR RESEARCH: THE CRAWFORD'S ET AL. CONTRIBUTION (2012, 2014, 2015)

Let us analyse more in detail Crawford's *et al.* work (2012, 2014, 2015). As we explained above, in the search for a comprehensive theory of emergence and growth that could explain new ventures performance, Crawford et al. (2012) asked: "*Are outcomes in the domain of entrepreneurship power-law distributed?"* (p. 4). They thought that the emergence of these power-law distributions could provide the foundation stone of a new theory, using the complexity science perspective. Taking into account those entrepreneurial outcomes that may be more relevant for constructing a theory of the emergence of firms and with more potential for practical considerations, Crawford and McKelvey (2012) selected six outcomes:

- Revenue.
- Number of employees.
- Revenue growth (%).
- Revenue gain (in absolute monetary terms).
- Number of employees' growth (%).
- Employee gain (in absolute numbers).

They considered both revenue and number of employees the most relevant outcomes for theoretical and practical purposes. They also included relative and absolute growth in revenue and number of employees, measuring the relative growth as percent (they use the term "growth" for the relative measure) and the absolute growth as difference in amount (they use the term "gain" for the absolute increase in amount).

Crawford and McKelvey (2012) hypothesized that these major entrepreneurial outcomes of emerging firms are power-law distributed. They analysed the outcomes of three samples in datasets from United of State of America:

1) Data collected in the Panel Study of Entrepreneurial Dynamics II (PSED II) that focused on the nascent entrepreneurial population (1214 subjects).
2) The Kauffman Firm Survey (KFS) (On-going businesses).
3) The Inc. 500® (INC) Extreme outcomes. Fastest-growing in USA. 500 companies with the highest growth rate published in Inc. Magazine.

Using MATLAB software, and the scripts, protocols and techniques for calculating power-law model fit developed by Causet, Shalizi and Newman (2009), the authors estimate a) the parameters for the slope α - the scaling exponent -, b) the minimum value in the distribution that shows power-law behaviour (x_{min}), c) the standard errors of the estimates and d) the goodness-of-fit (Kolmogorov-Smirnov (KS) tests).

Crawford and McKelvey (2012) found that – except one - all the models that they run supported their power-law hypotheses in the different datasets. Analysing the data distributions, they argued that the x_{min}, defined in the statistical procedure, identifies a tipping point, a threshold, in which the system – the entrepreneurial emergence - goes from an additive, linear state into a non-linear state. This point (x_{min}) - the minimum value in the distribution that shows power-law behaviour - separates the Gaussian and Paretian regions, the dotted line in figure 1b, the “Gaussian world” from the “Paretian world”.

a) Rank/Frequency Distribution on Linear Scales;
b) Distribution Plotted on Log-Log Scales

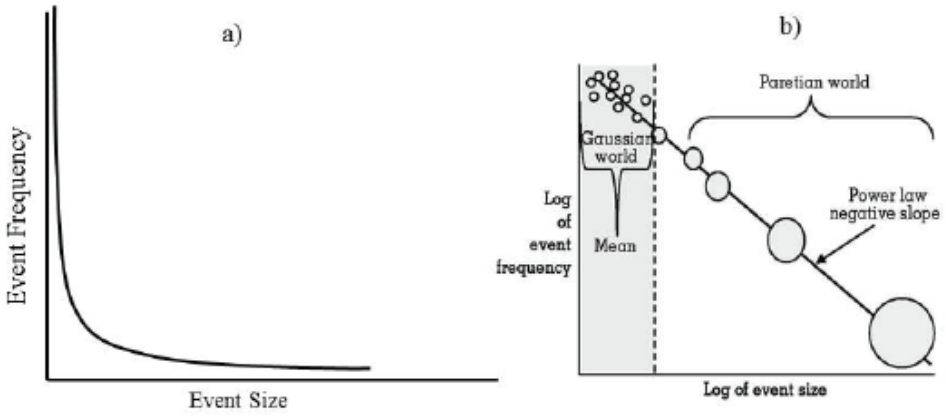


FIGURE 3 - ***FIGURE FROM CRAWFORD AND MCKELVEY (2012, FIG 1, P. 14).***

This point is called in complexity science the *threshold*, the *bifurcation point*, the *critical value*, the *phase transition point*, and, beyond this threshold, the system changes to a non-linear state, and firms operate "*in a much more interdependent, highly scalable, non-linear environment and, thus, have the potential to influence outcomes at a higher level*" (Crawford and McKelvey, 2012, p. 8; Lamberson and Page, 2012). It is the *region of emergent complexity*, where "*organisms are more likely to survive because they have a solid enough foundation of resources, yet maintain enough flexibility to change when environmental perturbations dictate*" (Crawford and McKelvey, 2012, p. 8).

According to Crawford and McKelvey (2012), firms, in the non-linear zone, beyond the tipping points, have the potential to influence greatly their environment producing non-linear outcomes, positive extreme events, and co-evolutionary effects. In the "*region of emergent complexity"*, as they call it, around the threshold or beyond the tipping point, firms have solid foundations and the required flexibility to adapt, to change and to success.

On the other hand, beyond the tipping point, unexpected negative extreme events can also occur if a firm has not compensated outcomes. The authors gave the example of a nascent firm with a non-linear number of employees but only linear revenue: would this firm survive having not compensated outcomes? Would therefore these power laws have predictive potential? Although Crawford and McKelvey (2012) did suggest it, and they proposed to use the tipping points as benchmarks to increase the probabilities of survival and promote growth, they did not go deeper into this issue in that paper. Furthermore, Crawford and McKelvey (2012) proposed several practical applications that could be derived from the study of these power laws such as in counselling, pedagogy, policy interventions, etc.

In Crawford et al. (2015) they extended the research done in 2012 and 2014 1) geographically, studying also the Australian panel data set, and 2) conceptually, incorporating not only **outcome** variables (revenues, employees) but also **input** variables, such as human capital resources variables, financial capital resources variables, cognitive variables (expectations), start up activities, and industrial sectors aspects (business environment). Again, results revealed that 48 out of 49 essential variables of more than 12,000 nascent, young, and hyper-growth firms in U.S.A. and Australia exhibited power law – heavy-tailed - distributions.

1.3.1 Predictive Potential Of Power Law Distributions

Crawford and McKelvey (2012) suggested several practical implications derived from their discovery, such as the predictive possibilities associated to power laws and aspects related to policy implementations. The forecasting potential of power laws and their linked complex systems has already been explored, for example, in geophysics and seismology

(Rundle et al., 2003). Based on the Gutenberg-Richter empirical law, a power law that formulate the relationship between the size (magnitude) of the earthquakes and their frequency (Gutenberg and Richter, 1954), it has been possible to develop methods for earthquake forecasting: the frequency of big earthquakes can be extrapolated from the frequency of the small ones (Sornette and Sornette, 1989).

Crawford & McKelvey (2012) briefly raised an example of a potential forecasting possibility of these power laws using the tipping points of these distributions. They considered the x_{min} - the minimum value in the distribution that shows power-law behaviour - as the critical threshold of the distribution: beyond this tipping point appears the region of *emergent complexity*.

Let us consider the results from the PSED for the fifth year:

Outcomes	n	med	mean	skew	s.d.	max	x_{min}	α	K-S
Employees									
PSED - Yr5	68	0	30	12	122	1,500	2 ± 1	1.73 ± 0.15	**0.06**
Revenue ($000)									
PSED - Yr5	145	42	35	7	1,400	12,000	600 ± 36	1.78 ± 0.12	**0.06**

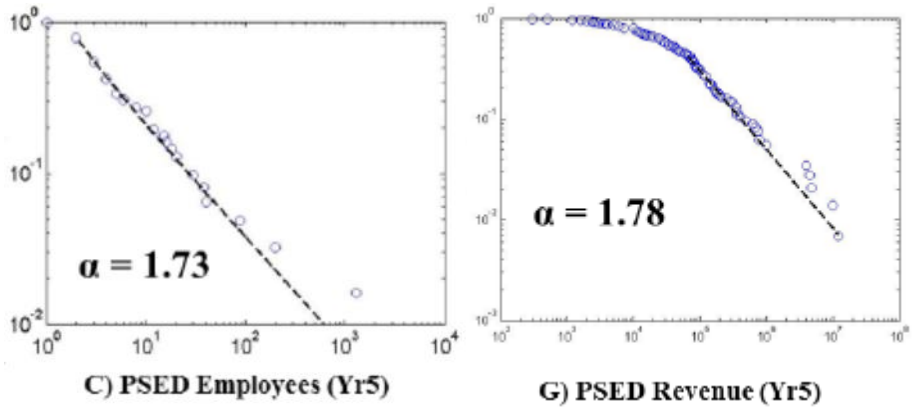


FIGURE 4 - FIGURE AND TABLE FROM CRAWFORD & MCKELVEY (2012, P. 9)

From the table and figures above, a firm is in the area of complexity (beyond the threshold x_{min}) when the number of employees in more than 3 (2 ± 1), and the revenue is beyond $600,000. Now, what would happen if a firm has a number of employees in the complexity zone, let us say 4 employees, and, however, only a linear revenue, for example, the mean revenue value in the fifth year ($35,000)? Crawford and McKelvey (2012) suggested that a negative extreme event such as lay-offs or firm closure might occur because the firm has not compensated outcomes: it may be financially difficult to support 4 employees with a linear revenue. A more adequate revenue amount, beyond the complexity threshold of $600,000, could allow the firm to survive. Crawford and McKelvey (2012) did not enter in more detail in this paper. It would be necessary to analyse the data set, firm by firm, to check the accuracy of this statement. Would firms with not compensated outcomes – values in different regions of the thresholds - collapse? This type of analysis also would be applied to other similar datasets such as the Australian CAUSEE or the UK PSED (Reynolds and Curtin, 2011; Reynolds, Hart and Mickiewicz, 2014).

There would also be policy and practical implications. Should a public institution trying to promote entrepreneurship give a grant to a nascent firm that has not compensated outcomes? If this study concludes that that firm would collapse because lack of compensation, would this be a

waste of tax payers' resources? Would a venture capital firm invest in an uncompensated company? Would this investment fail? Or should a venture capitalist focus its investments on those areas of the firm (outcomes) that are not beyond the complexity threshold – in the area of the power law - in order to help them to compensate the outcomes and allow the firm to survive and flourish?

1.4. NEW DEVELOPMENTS: NEW STATISTICAL METHODS AND SOFTWARE PACKAGES AND THEIR IMPACT ON THE STUDY OF HEAVY-TAILED DISTRIBUTIONS IN ENTREPRENEURIAL DATASETS.

Crawford et al.'s (2014, 2015) based their statistical analysis in the method developed by Clauset, Shalizi and Newman (2009) that combines maximum-likelihood fitting techniques with goodness-of-fit tests based on the Kolmogorov-Smirnov statistic and likelihood ratios (also in: Virkar and Clauset, 2014). Several new implementations of the methods described in Clauset et al. (2009) article have been proposed since then, that have made it much easier to evaluate the best fit among different alternative distributions (pure power law, power law with cut-off, exponential, log-normal, etc.) (Alstott et al., 2014 for Python, Gillespie, 2015 for R, 'plpva.m' function for Matlab, C++, etc.).

The application of these new developments in statistical software packages, such as the **poweRlaw** package for R (Gillespie, 2015), led Shim (2016) to the conclusion that lognormal distributions – another kind of heavy-tailed distributions, rather than pure power laws - were a better fit for entrepreneurial outcomes. Shim (2016) applied Clauset et al.'s (2009) fitting

techniques to several variables from the Panel Studies of Entrepreneurial Dynamics II (PSED II) in USA (Reynolds and Curtin, 2008) checking not only power law distributions but also other alternative models such as log-normal and exponential distributions. **He found that lognormal distributions were the best model for these American entrepreneurial outcomes variables** and that the distributions change into power law over time. That is, in the early stages of the new ventures, the outcome variables follow a log-normal pattern and, after the emergence of the venture, the outcome variables turn into power law distributions. Shim proposed then a transitional process from lognormal to power law. The ventures' early-stage outcome distributions are less skewed; over time, those distributions will change to more skewed power law patterns (Shim, 2016). This proposed transition from log-normal distribution to power law in nascent entrepreneurial outcome datasets made perfect sense because it is possible to observe similar transitions in nature and social sciences. For example, at the beginning of the dataset, the distribution of income shows lognormal distribution with the lower/medium income, but, with large incomes, the dataset becomes an inverse power law, the Pareto's law. Similarly, this transition can also be observed in the distribution of the number of papers published by scientists, the Lotka's Law (West and Deering, 1995).

Based on Mitzenmacher (2004) and on Nirei and Souma (2007), Shim (2016) also suggested a multiplicative process as the possible generative mechanism of these long-tail distributions and he performed a simulation in **R** software to determine whether this hypothesis was plausible. He found that the log-normal distribution was a better fit than the power law model at every stage of the simulation, unlike the empirical results. Thus, his computer simulation results cast serious doubts on the theory of the outcome distribution change over time – from log-normal to power law - mentioned above.

Continuing also the path opened by Andriani and McKelvey (2009), recently, Joo, Aguinis and Bradley (2017), using a novel distribution pitting technique – a new fitting software - have found that the exponential tail distributions - exponential and power law with an exponential cut-off - and their generative mechanism - incremental differentiation -, are the most frequent distribution in many individual outputs across different organizations, sectors, jobs and activities. However, as Shim (2006) pointed out, this may not be totally accurate in nascent entrepreneurship processes: **this thesis research will show that the lognormal distribution in entrepreneurial outcomes seems predominant throughout the different panel studies in different countries**. The power law distribution with an exponential cut-off may also be a plausible fit in some particular panel outcomes variables.

Applying complexity science methods and tools in the field of entrepreneurship, this study will continue the search for the identification of heavy tailed distributions in **nascent entrepreneurial longitudinal datasets** in different countries, and it will explore the processes that origin the emergence of these kinds of distributions. This research will focus on the period of time that elapses before becoming an established firm, i.e. before being a fully established organization, in the period in which **nascent entrepreneurs** carry out the decisive decisions and actions that would lead to venture emergence. This study is not about “new” firms, but rather about “emerging” firms, **nascent ventures** in the process of becoming, nascent entrepreneurial processes, and nascent entrepreneurs. This period of time centred on nascent entrepreneurs and the process of organization creation has also been called “organizational emergence” (Gartner, Bird and Starr, 1992), the “preorganization” (Katz and Gartner, 1988; Hansen, 1990), “gestation” – using the biological metaphor (Reynolds and Miller, 1992) - , or start-up (Carter, Gartner and Reynolds, 1996), and it is the period of time targeted by longitudinal panels on entrepreneurial activities such as the U.S. Panel Studies of Entrepreneurial Dynamics (PSED) and their counterparts in other countries.

1.5. THE METHODOLOGICAL SHIFT: MODELLING COMPLEX SYSTEM WITH NEW MODELLING TECHNIQUES. THE AGENT-BASED APPROACH

Agent-based modelling and simulation (ABMS) is a relatively recent methodology for modelling complex systems, related to the research of non-linear dynamics and artificial intelligence, based on interacting, autonomous 'agents' that was facilitated by the arrival of personal computers in the 1980s and early 1990s.

> *"An agent-based model is a computer program that creates an artificial world of heterogeneous agents and enables investigation into how interactions between these agents, and between agents and other factors such as time and space, add up to form the patterns seen in the real world" (Hamill and Gilbert, 2016, p. 4).*

This ***agent perspective*** is the most essential and distinctive characteristic of ABMS: the system is viewed as made of agents in interaction with other agents and the environment (Macal, 2016). Agents behave according to rules and they interact with other agents. These agents also interact in space and time according to rules. These behaviours and interactions of the agents at a micro-level may produce a distinct behaviour of the system as a whole. New patterns and structures may emerge, without explicit previous programming into the model, that arise by the combination of agents' attributes, behaviours and interactions (Macal and North, 2010). These interactions at a micro-level, the aggregation of these micro-level and meso-level behaviours, may create emergent patterns at a macro-level: these patterns emerge from the bottom up (Page, 2008; Schelling, 2006, prev. ed. 1978).

A standard agent-based model has three basic features:

- A set of **agents**, with their attributes (state variables) and behaviours.
- A set of agents' methods of **interaction** ("rules of engagement"). There is an underlying topology of interconnection that defines how and with whom agents interact.
- The agents' **environment**: The environment affects agents and their interactions.

Agents have the capability to act autonomously, that is, to act by themselves without external direction depending of the situation. Agents have a set of rules and behaviours that allow them to take independent decisions.

From the practical modelling perspective, agents have the following characteristics:

- An agent is a self-contained and a uniquely identifiable individual. It has a boundary. In addition, it has attributes (state variables) that make that agent different from other agents.
- Agents are autonomous. They are independent based in the environment and in the interaction with other agents. The behaviour of an agent can be defined by simple rules or sophisticate adaptive mechanisms such as neural networks or genetic algorithms.
- An agent has state variables that change over time. These state variables are related to the agent's attributes.
- An agent has a dynamic interaction with other agents that affect its behaviour. An agent has a set of protocols for

interaction with other agents: communication, how to move and respect the topology of the model world, how to respond to environment, etc.

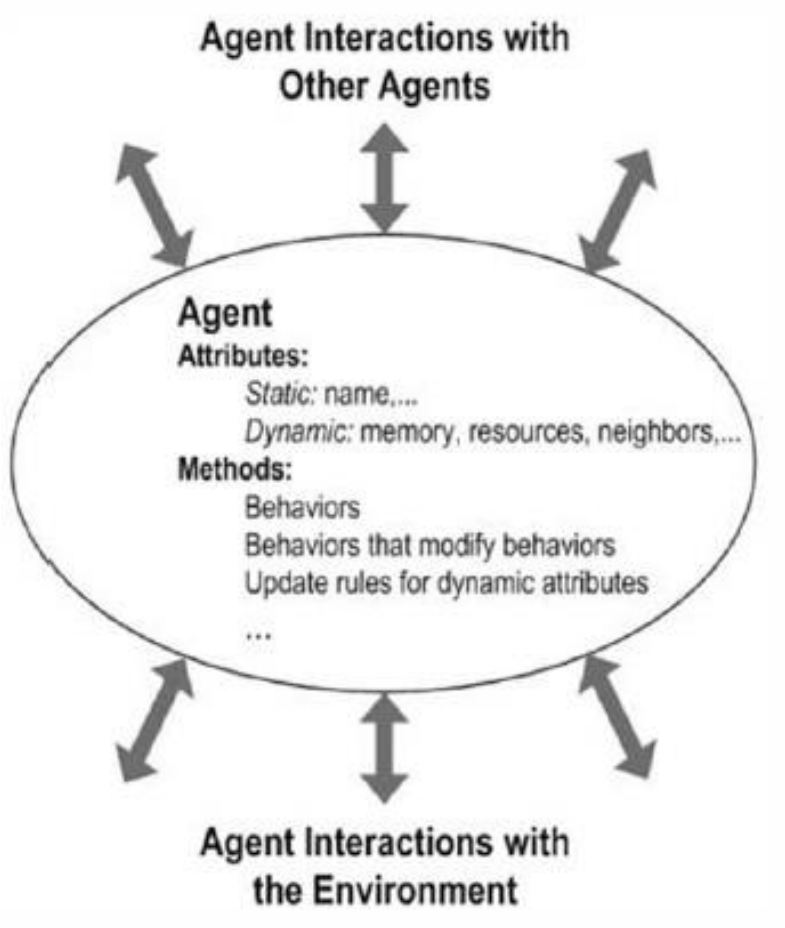


FIGURE 5 - A TYPICAL AGENT ACCORDING TO MACAL & M J NORTH (2010, P. 154, FIGURE 2)

ABMS (Agent-Based Modelling and Simulation) draws its theory and concepts from complexity science, systems science, computer sciences and artificial life (Macal and North, 2009). The study of Complex Adaptive Systems (CAS) is the historical root of ABMS, in which systems are also built from the ground-up (Kauffman, 1993; Holland, 1995). Complex Adaptive Systems (CAS) address the question of how complexity arises from autonomous agents, and it was initially focused on adaptation and

emergence of biological systems. Complex Adaptive Systems (CAS) are those that can self-organize and dynamically reorganize to be able to survive in their environments (adaptive ability): they are defined by a group of interacting agents, who can act and react to the actions of the other agents. Examples of CAS are ecosystems, financial markets, colonies of ants, etc. Emergence, defined as a macro or meso-level phenomenon as a result of local micro-level interactions, is one of the most important phenomena that can occur in Complex Adaptive Systems (Macal and North, 2009). ABMS were fundamentally developed as the corpus of ideas, techniques, and tools for implementing computational models of complex adaptive systems (Macal and North, 2010). ABMS (Agent-Based Modelling and Simulation) can be used both for the investigation of the dynamic of a process – a simulation - or for developing models designed to do optimization, such as particle swarm optimization and ant optimization algorithms.

The application of Agent-Based Modelling and Simulation (ABMS) across natural, social and physical disciplines is growing unceasingly, despite the debate on its methodological nature and current discussions on their proper implementation and development (Grimm et al., 2014; Grimm and Berger, 2016; Macal, 2016). Since the publication of the book *Growing Artificial Societies* (Epstein and Axtell, 1996), it has been a continuous development of new agent-based models with diverse applications, new methods and theory building.

Agent-Based Modelling and Simulation (ABMS) has already been used to address many complex systems phenomena both in natural processes (Vicsek, 2002), in social sciences (Bonabeau, 2002b, Epstein 2006, Gilbert, 2008), economics (Farmer and Foley, 2009; Hamill and Gilbert, 2016) and management (Davis et al., 2007). New computational capabilities have made possible to apply these ABMS techniques to different disciplines and subjects, from modelling agent behaviour in stock

markets (Arthur et al., 1997) to simulating and predicting the spread of an epidemical disease (Macal and North, 2009). Currently, applications of agent-based model can be found almost in all disciplines not only in the natural, social or physical sciences but also in engineering, business, operation management, and similar fields, becoming a common simulation technique (Fioretti, 2013; Macal, 2016). However, there are discussions about the nature of ABMS, how to develop the models and the relationship with other types of simulation and modelling because ABMS is used in many, different scientific communities, each of them with a different interpretation. Given that ABMS can offer an explicit framework for modelling people (agents) behaviours, social interactions and social processes, it has become the leading method to model societies and organizations (Robertson and Caldart, 2008). Besides the traditional inductive or deductive methods, the possibility of grow artificial societies using ABMS opens a new kind of *generative* social science (Epstein, 2006), a third way of doing science (Axelrod, 1997). One of the main reasons to develop ABMS is because this modelling technique allows a better representation of human behaviour and the discovery of the collective effects of organizations and societies. Fields such as behavioural economics or behavioural operation management are looking for improved models of behaviour, in which bounded rational agent model can be introduced including realistic constrains on time, effort, information, capabilities, etc. (Simon, 1991; Balke and Gilbert, 2014; Macal, 2016). Epstein (2014) has even proposed the possibility of incorporating neuroscience knowledge into ABMS to replicate emotions, cognitive and social aspects of agents. Agent-based modelling can be a useful tool for incorporating neuroscience theory and methods into entrepreneurship research (de Holan, 2014; Nicolaou and Shane, 2014).

1.6. AGENT-BASED MODELLING AND SIMULATION IN SOCIAL SCIENCES AND ENTREPRENEURSHIP

McMullen and Dimov (2013) – and prior to them, McKelvey (2004) - have proposed that Agent-Based Modelling and Simulation (ABMS) will be a very important tool and methodology for the generation of theory in entrepreneurship. In spite of the fact that the number of management and operations researchers interested in computer simulation methods has increased in the last 20 years (Davis et al., 2007, 2009; Harrison et al., 2007; Robertson and Caldart, 2008; Günther et al., 2011), agent-based simulations remain scarce in entrepreneurship research (Aldrich, 2001; Coviello and Jones, 2004; Van de Ven and Engleman, 2004; Yang and Chandra, 2013). Entrepreneurship scholars have been slower to adopt agent-based modelling (McDonald et al., 2015), compared with those from natural science and other social sciences such as economics (Tesfatsion, 2002) or sociology (Sawyer, 2003). The publication of papers on entrepreneurship using agent-based models has just started few years ago (Bhawe et., 2016; Shim, Bliemel, & Choi, 2017; Breig, Coblenz & Pelz, 2018).

Social agent-based modelling, that is, to model social processes from the individual level, “from the ground up”, has been developed since 1970s, using, for example, cellular automata models (North and Macal 2007). Epstein and Axtell have suggested several social processes that could be successfully agent-based modelled (Epstein and Axtell, 1996; Epstein, 1999; Epstein, 2006). Together with Schelling's segregation model (Schelling, 1969), the *Sugarscape* model of Epstein and Axtell (1996) have been the most well-known agent-based models in social sciences. This bottom-up computational modelling can be readily applied to entrepreneurship considering the “entrepreneur” as an agent – with its attributes - that interacts in a complex way with other agents - of similar or different nature - and environments. On the other hand, entrepreneurs fulfil the major characteristics of the agents in ABMS: they are autonomous,

interdependent, and adaptive, and they follow operational, behavioural and strategic rules (McMullen and Dimov, 2013; Miller and Page, 2007).

Yang and Chandra (2013) in their paper “"*Growing artificial entrepreneurs: Advancing entrepreneurship research using agent-based simulation approach*" offered a rationale of this methodology for entrepreneurship research, and sketched a roadmap for its use in this field (Yang and Chandra, 2013):

> *“(…) agent-based simulation approach can be useful for explaining, discovering – and thus formulating formal theory – and predicting the unpredictable phenomena in entrepreneurship. (Yang and Chandra, 2013, p 227).*

Yang and Chandra (2013) argued that there are shared conceptual foundations between entrepreneurship and ABMS such as autonomy, heterogeneity, bounded rationality, learning, and disequilibrium. Yang and Chandra (2013) examined the possibilities of formalizing the entrepreneurial processes into ABMS code based on empirical facts and generally accepted foundations of entrepreneurship, and they considered that computer simulations can advance entrepreneurship research because those models allow the analysis of internal validity of theories of entrepreneurship and can be explored through “systematic experimentation”. Computer-simulations, as part of the “science of the artificial” (Simon, 1996; Sarasvathy, 2003), allow the researchers to test the robustness of their theories on entrepreneurship and to understand, explain and predict the implications of those theories. These tasks may result very difficult using other research methodologies (Gilbert and Terna, 2000). Experimentation in ABMS is implemented changing the rules of behaviour or introducing new agents, or varying different scenarios to discover their impact on the global system (Yang and Chandra, 2013). Our model “*A nascent entrepreneurial agent-based model*” is partially rooted in the

definitions of the rules and assumptions about agents from the conceptual model theoretically described by Yang and Chandra (2013).

As we mentioned above, one of the most recent attempts of the simulation of entrepreneurial outcomes distributions was initially developed by Shim (2016) using **R** software. He performed a simulation to determine if heavy-tailed distributions can be obtained through multiplicative processes in entrepreneurship. Shim (2016) was able to show that the distributions of the simulated outcomes were quite similar to the empirical datasets and that lognormal models have better fit than other heavy-tailed distributions in most of the nascent venture early stages (activities) results. However, Shim (2016) suggested that more sophisticated agent-based modelling and simulations were needed, given that a random multiplicative process was not enough to explain the complexity of the empirical and simulated patterns.

Based on a bibliometric method and on the behavioural rules inferred from the entrepreneurship literature, Shim, Bliemel and Choi (2017) proposed a simple agent-based model based on essential concepts – “stylized facts” -, that was able to simulate the emergence of heavy-tailed distributions in nascent venture outcomes and that was consistent with the empirical datasets. Their model consists in two agents (“entrepreneur” and “investor”) and two objects (“opportunity” and “resources”), being the amount of resources modelled as state variables of entrepreneurs and investor. Breig, Coblenz and Pelz (2018) has recently proposed another simulation model, used as an illustrative example of statistical validation for the entrepreneurial variable “venture debt” with the empirical data extracted from the second Panel Study of Entrepreneurial Dynamic (PSED II).

However, in order to explore more complex phenomena in nascent entrepreneurship or to introduce other essential elements of this nascent

entrepreneurial process, a more complex agent-based model would be required. **Our objective is to introduce a baseline model with new layers of complexity to previous entrepreneurial agent-based model attempts.** Our model is designed to explore questions regarding the emergence of new ventures and their nascent entrepreneurial processes, and to identify the mechanisms that produce the emergence of heavy-tailed distributed outcomes ("patterns", Grimm, 2005) in nascent entrepreneurs' longitudinal data panels (PSED and similar empirical datasets). Although our model adopts most of the basic features and conceptual framework used in previous models (especially the conceptual model of Gartner, 1985) and the roadmap proposed by Yang and Chandra (2013), it introduces new levels of complexity in comparison to them (Shim, 2016; Shim, Bliemel and Choi, 2017; Breig, Coblenz and Pelz, 2018). Our model has additional features, more internal state variables for agents, additional forms of interactions among them, additional rules of behaviour and types of agents, new global environmental variables (Martinez, Yang and Aldrich, 2011), that allow the possibility of further research in relationship with the empirical data: calibration, parametrization, verification, etc. One of the purposes of this model is to expand previous "stylized fact" type of agent-based modelling - based on basic principles - to richer representation of real-world scenarios based on empirical datasets. A more complex model also allows deeper theory development from simulation (Davis et al., 2007).

Our model starts with the discovery of the heavy tailed distribution patterns at the macro level – the "stylized fact" -, and it tries to simulate the underlying processes and behaviours of individual entrepreneurs at the micro level that produce that "stylized fact" (the pattern: the heavy tailed distribution) (Shim, Bliemel and Choi, 2017).

1.7. RESEARCH OBJECTIVES

This research has two major sections:

1) **Data analysis**: Extension of the empirical datasets analysis of other international longitudinal panels freely available and exploration of their distribution patterns (Crawford et al., 2012; Crawford et al., 2015; Shim, 2016; Shim et al., 2017).

2) **Design and implementation of an agent-based model as a research tool,** with enough complexity to be able to simulate the heavy tailed distributions patterns in the different international empirical longitudinal studies, and as a baseline research tool - openly available to the research community - to test and explore new theories and empirical datasets in nascent entrepreneurial processes.

DATA ANALYSIS

Crawford and McKelvey (2012) discovered ubiquitous power law distributions in the Panel Study of Entrepreneurial Dynamics II (PSED) that assesses the level of initial entrepreneurial activity in a representative sample of American nascent entrepreneurial population (Reynolds and Curtin, 2008). The Panel Studies of Entrepreneurial Dynamics (PSED) has also been conducted in other countries using similar methodology over the last 15 years: Canada, Netherlands, Norway, Sweden, United States, Australia, China, Germany, Latvia, and UK (Reynolds and Curtin, 2011; Reynolds, Hart and Mickiewicz, 2014). Crawford et al. (2015) also identified power law distributions in outcome variables in the Australian longitudinal study (CAUSEE).

Research questions:

- Are these heavy-tailed distributions also present in similar longitudinal studies conducted in other countries?
- Which heavy-tailed distributions (power law, log-normal, etc.) are the best fit for these data?
- What are the generative mechanisms that produce these heavy-tailed distributions?

THEORY DEVELOPMENT THROUGH AGENT-BASED MODELLING AND SIMULATION (ABMS).

Agent-Based Modelling and Simulation (ABMS) has already been used to address many complex systems phenomena both in natural processes (Vicsek, 2002) and in social sciences (Epstein and Axtell, 1996; Cederman, 2005; Epstein, 2006), and it has proven its capacity to generate theory, "agent-based generative theory" (Epstein, 1999; Davis et al., 2007). Given the characteristics of the phenomenon of emergence of heavy-tailed distribution in entrepreneurial processes, originated in the interactions of multiple agents in a specific set of conditions, it seems reasonable to use agent-based modelling and simulation techniques to model this emergence.

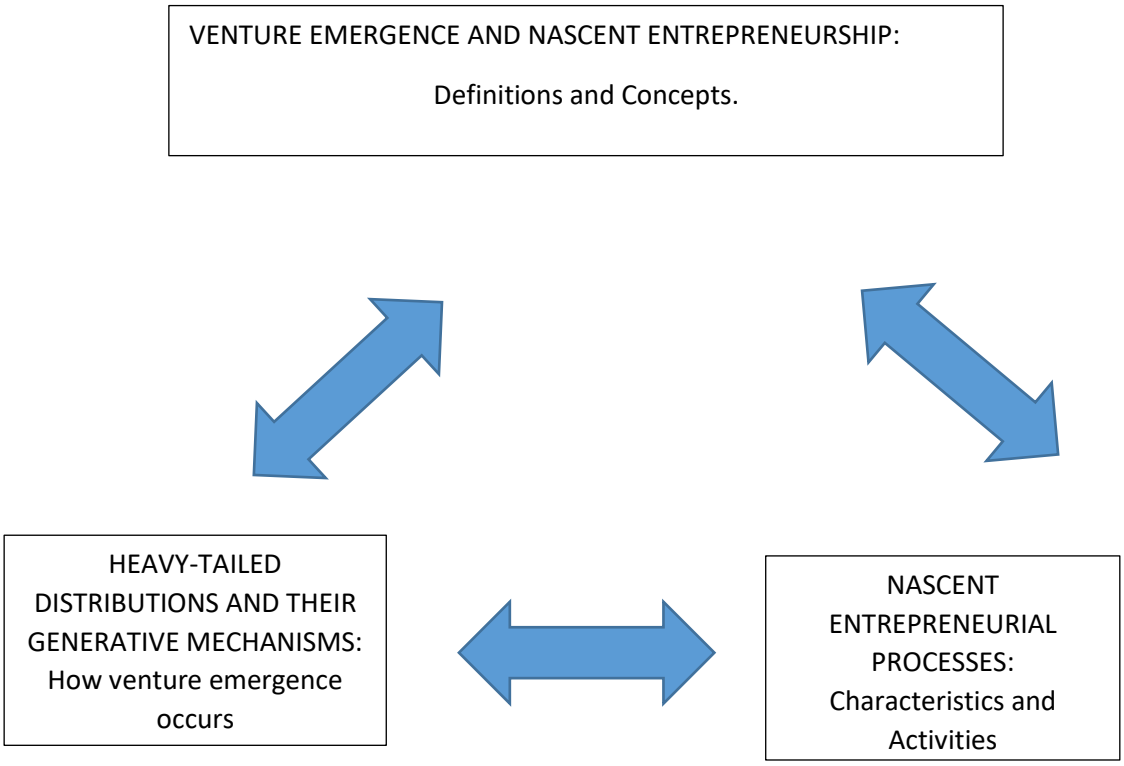
Research questions:

- How can the mechanisms that produce the emergence of heavy-tailed distributed outcomes in nascent entrepreneurial processes be simulated by an agent-based model with a complex set of agents' variables and behaviours, and be parametrized and calibrated with empirical datasets?

- How can a versatile baseline model be designed and implemented to explore other international longitudinal panel datasets?

2. LITERATURE REVIEW

2.1. VENTURE EMERGENCE AND NASCENT ENTREPRENEURSHIP: DEFINITIONS AND BACKGROUND.



2.1.1. "EMERGING" FIRMS VERSUS "NEW" FIRMS: THE CONCEPTS OF VENTURE EMERGENCE AND NASCENT ENTREPRENEURSHIP.

Heavy-tailed distributions have been reported on **already established firms**. The literature on already established firm size distributions and industrial organization population dynamics will be explored in the next section under the title "Heavy tail distributions in

Economics: Antecedents" and it will provide a relevant background and useful insights in the modelling section of this thesis.

This research, however, focuses on the emergence of new ventures and the nascent entrepreneurial processes related to them: it deals with the series of events that happen **before becoming a firm**. It analyses the processes that occur before of being an established organization, in that period in which nascent entrepreneurs carry out the decisive decisions and actions that would lead to venture emergence (Reynolds, 2017). As Carter, Gartner and Reynolds (1996) have pointed out, to study new organizations is not the same that to study emerging organizations (Ács and Audretsch 2010), given that the activities and processes related to maintaining or modifying the operations of established firms are not the same that those related to the creation of new organizations (Gartner et al., 2010).

The distinction between "new organizations" and "emerging organizations" is methodologically decisive in this research. Studies on entrepreneurs who are operating already established new businesses provide partial information about the process of organization creation: it assumes the outcomes of emergence – the established firm - without providing information regarding those entrepreneurs that tried to create a new organization and failed (Gartner et al., 2010). To study only entrepreneurs who have successfully started a new venture introduces a selection bias, with no information on start-up activities on nascent entrepreneurs that failed in their attempts (Delmar and Shane, 2004).

Assuming that a "new" venture is not the same as an "emergent" venture, this research will focus on studies that use samples of nascent entrepreneurs, that is, it will be centred in those studies that analyse what happens in the process of starting a business rather than on those studies that survey entrepreneurs of new on-going firms (Gartner et al., 2010).

Therefore, here, a **nascent entrepreneur** - the subject of our research - is defined as someone in the process of establishing a new venture but who had not yet succeeded in making the transition to new business ownership (Carter et al., 1996; Dimov, 2010).

A nascent entrepreneur seeks a business opportunity, that is:

- to introduce a new product or service, or
- to open a new market, or
- to develop a more efficient and profitable production method (Shane and Venkataraman, 2000).

A **nascent or emerging venture** is considered the sum of the efforts, actions and judgments carry out by the nascent entrepreneur. At some point in time, an emerging venture may become a new venture, or be extinguished – or even remain latent -. During the process, the emerging venture will receive increasing inputs, not only by the nascent entrepreneur – who is essential at the beginning - but also from other stakeholders, such as new partners, resources, financial institutions, etc. (Dimov, 2010).

New data on the process of starting new ventures and the nascent entrepreneurs' activities have been provided by surveys such as the Panel Studies of Entrepreneurial Dynamics (PSED) – PSED I and PSED II -. Before the existence of the PSED studies, literature on this emerging period was scarce. Most of the published entrepreneurship research was based on samples of already established and existing firms. Studies on the earliest phases, before becoming a firm, for example, in Carter, Gartner and Reynolds (1996), were rare (Davidsson and Honig, 2003).

The Panel Studies of Entrepreneurial Dynamics – PSED I and PSED II- were detailed longitudinal surveys that were able to identify a representative sample of nascent entrepreneurs in United States, and have

generated important information into the process of how ventures emerge. In PSED, a nascent entrepreneur is identified and classified as such if this person initiated at least one start-up activity by the time of the interviews, among a number of other potential entrepreneurial gestation behaviours (see below a list of these gestation activities).

2.1.2. Longitudinal Methods: The Entrepreneurial Process as a Dynamical Phenomenon.

This research will make use of longitudinal panel studies such as the PSED II responding to the call for considering **entrepreneurial activity as a process** rather than a punctual act, and taking into account the role of time in this phenomenon (McMullen and Dimov, 2013, p. 1482):

> *"Prior work has thus tended to diminish the role of time in the entrepreneurial process by studying entrepreneurship as an act, as opposed to a journey that explicitly transpires over time. To look forward, we reiterate and illustrate the tenets of a process approach by paying attention to the unit of explanation, logic of causal relationship, and nature of cause. We propose that a shift in inquiry from act to journey may advance scholarly understanding of the entrepreneurial phenomenon by evoking a number of challenging questions (McMullen and Dimov, 2013, p. 1482).*

The processes of organization formation have to be considered, therefore, a fundamental core of entrepreneurship (Gartner, 1985; Carter et al,, 1996; Gartner et al., 2010).

PSED II, started in 2005 as an improved replication of PSED I, provides a description of the initial stages of the entrepreneurial process. It makes a series of follow-up interviews of an initial cohort of 1,214 nascent entrepreneurs (Reynolds and Curtin, 2008). Longitudinal studies on venture creation, such as the Panel Studies of Entrepreneurial Dynamics (PSED), are able to identify those individuals entering in the start-up of the new firms

and to follow up their activities and outcomes during several years. They tracked the development of new ventures, from the emergence of a business idea and the organization of the start-up team, through the birth of an operational and legal registered firm.

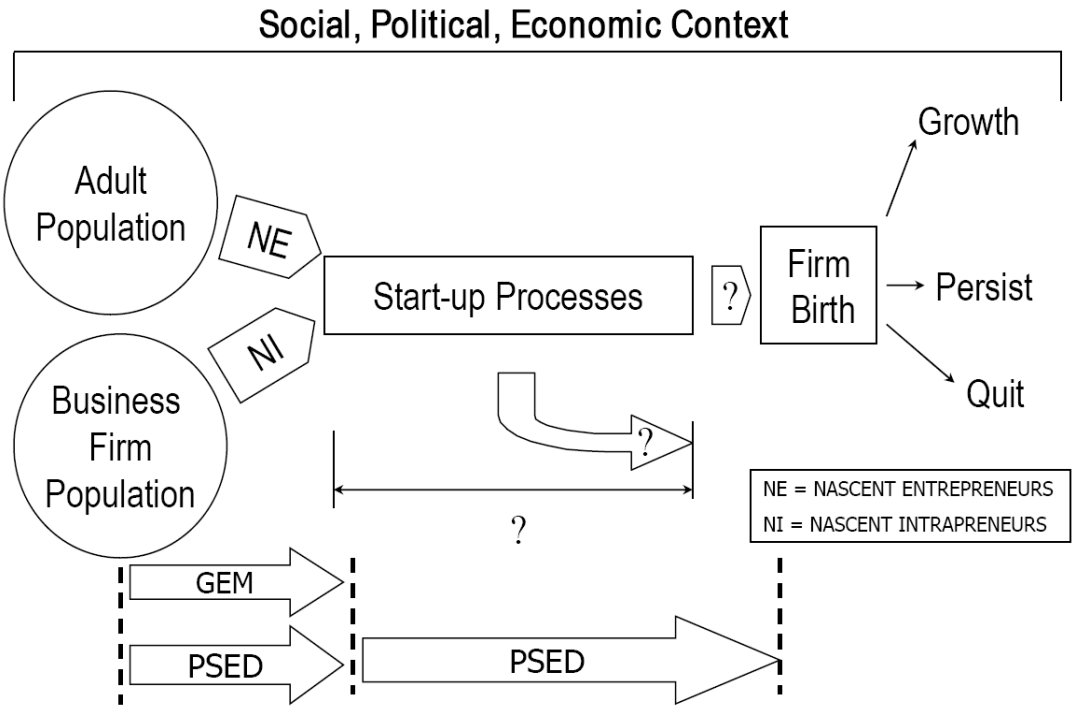


FIGURE 6 - THE QUESTION MARKS IN THE FIGURE POINTS OUT THE PERIOD TO BE STUDIED IN THIS RESEARCH: IT IS WHEN A NASCENT ENTREPRENEUR UNDERTAKES THE PROCESS OF ORGANIZATION CREATION ("START-UP PERIOD", "ORGANIZATIONAL EMERGENCE", "GESTATION"). FIGURE FROM CURTIN AND REYNOLDS (2007), P.12.

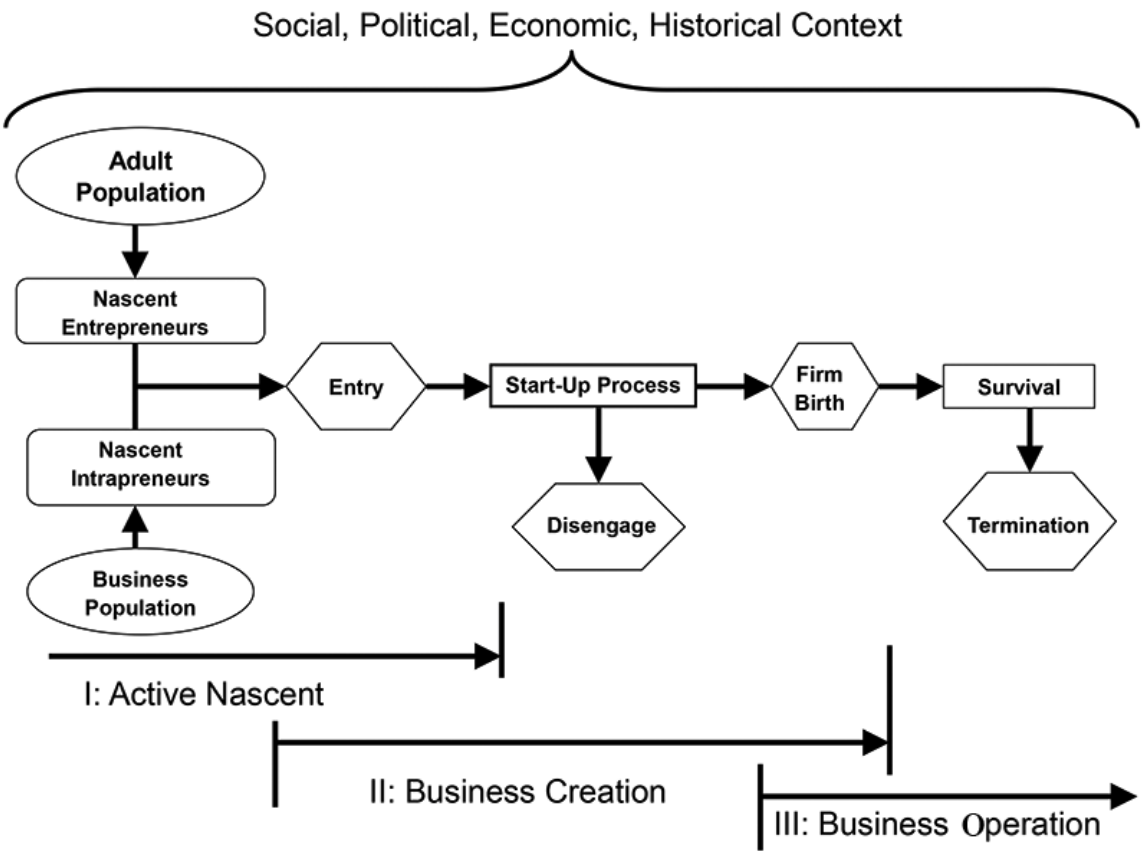


FIGURE 7 - CONCEPTUALIZATION OF THE ENTREPRENEURIAL PROCESS . FIGURE FROM RAYNOLDS 2017B.

These panel studies provide information such as:

- The length of time required to start-up and to constitute new firms.
- The amount and types of activities before registering the new firm in the registries or Chambers of Commerce.
- The amount and type of financial resources - formal or informal - that is gathered before the new firm is registered.
- The strategies and business models that these nascent ventures implement.

- The nature, composition, and background of the entrepreneurial teams.
- The use of and reaction to entrepreneurship promoting programs.
- The proportion of start-up ventures that become profitable and viable new firms.
- The main aspects of the transition to a profitable and viable new firm or to disengagement.

The book *New Business Creation: An International Overview*, edited by Reynolds and Curtin (2010), makes an extensive analysis of these longitudinal studies in different countries: The U.S. projects (the first and second Panel Studies of Entrepreneurial Dynamics, PSED I and II) and their counterparts in a number of other countries such as Australia, Canada, China, Latvia, Netherlands (two projects), Norway, and Sweden. These projects have been implemented over the past decade, and they are at different stages of development. Currently, only the complete datasets of four of these projects are publicly available (Australia, Sweden, US PSED I & II):

- Australia: “The Comprehensive Australian Study of Entrepreneurial Emergence“ (CAUSEE).
- Canada: “The Canadian Panel Study of Entrepreneurial Dynamics”.
- China: “Anatomy of Business Creation in China: Initial Assessment of the Chinese Panel Study of Entrepreneurial Dynamics”.
- Germany: “German Panel of Nascent Entrepreneurs”.
- Latvia: “Panel Study of Entrepreneurial Dynamics Overview”.
- Netherlands: “New Business Creation in the Netherlands”.
- Norway: “Business Start-up Processes in Norway”.

- Sweden: “The Swedish PSED: Performance in the Nascent Venturing Process and Beyond”.
- United States: “Panel Study of Entrepreneurial Dynamics I, II”.
- The UK 2013 Panel Study of Entrepreneurial Dynamic has also been implemented (Reynolds, Hart and Mickievicz, 2014).

Eventually, the principal outcome of the process of entrepreneurial activity is that the organization comes to existence or not. Other outcomes may occur in the process: the creation of new products or services, new customers or segments of costumers, etc. However, the identification and the definition of whether and when there is a new organization is a challenge (Gartner et al., 2010; Reynolds, 2017a). Different measures have been used – e.g., sales, business license, etc. - but none of these measures are able to capture fully by themselves whether an organization exists. For example, an entrepreneur may have obtained a business license to operate and, however, he or she may not have a clear idea about what is the objective of the firm, he or she may not have any sales, or he or she may not have a physical location or any specific human or financial resources yet. On the other hand, although a first sale often signals a nascent firm’s eventual emergence, sometimes, it may happen very early in the process, when the emerging organization may be not fully operational or registered (Carter et al., 1996; Davidsson and Honig, 2003).

Katz and Gartner (1988) sought to identify a theoretical and empirically based framework for identifying the properties of emerging organizations. They found that many of the proposed theories in the literature assumed properties that happen only after organizations achieve some particular size, instead of some set of characteristics that can differentiate an emerging organization from other types of social situations (Carter et al., 1996). Katz and Gartner (1988) suggested four emergent

properties that would be indicators that an organization is in the process of coming into existence:

- **Intention**: activities that show purpose and goals, such as membership lists of entrepreneurial organizations, subscription lists to entrepreneurial magazines, client lists of specialized organizations in entrepreneurship (entrepreneurial training companies, etc.), membership to entrepreneurs networks, etc.
- **Resources**: search of human and financial capital, such as applications for loans from banks, savings and loans, finance companies, directories of new occupants in office buildings and commercial centres, venture capital proposals, etc.
- **Boundary**: conditions that distinguish the firm, such as a tax number, phone listings, licenses, permits, etc.
- **Exchange**: transactions between the emerging firm and others stakeholders, such as sales, loans, or investments, Chamber of Commerce membership, etc.

An emerging organization would flag itself in different ways, at different times, during the process of creation. Organizing is a process not a state (Katz and Gartner, 1988; Delmar and Shane, 2004). Katz and Gartner's (1988) properties are a way to explore the emergence of the organizations and to identify firms in the process of emergence. Given that entrepreneurship is a process and that the various properties of venture emergence appear over time, Gartner, Carter and Reynolds (2010) have proposed to consider the emergence of a firm as having sequential "birthdays", with these "birthdays" being the different measures used to identify a new organization (start-up team personal commitment, first sale, first employee, first outside financial support, etc.), although the sequence

of appearance of these properties seems to differ for the different industrial sectors (Reynolds and Miller, 1992).

2.1.3. The Processes Of New Venture Emergence: Characteristics and Activities.

Wiklund et al.'s (2011) defined entrepreneurship as an "organizational phenomenon" (Gartner et al., 2010, p 99), the phenomenon of "emergence of new economic activity":

> *"We strongly recommend that entrepreneurship research be unified as a field approached theoretically and empirically in terms of the phenomenon. We propose that the phenomenon of "emergence of new economic activity" lies at the heart of entrepreneurship (where "economic" has a much wider meaning than "commercial")." (Wiklund, 2011, p. 5).*

Davidsson states: "Entrepreneurship is about emergence" (Davidsson, 2003, p.55). Current definitions on entrepreneurship focus on the concept of **emergence**, suggesting that research should analyse the early phases of the phenomenon, the mechanism of detecting opportunities and how they are acted upon, or how new ventures appear (Gartner, 1988; Shane and Venkataraman, 2000). However, empirical knowledge on entrepreneurship using this emergence approach is still limited (Davidsson and Honig, 2003). This research deals with entrepreneurship using the methods and tools that are already in place for analysing other emergent phenomena that occur in nature and in other social context (Goldstein, 2011). Therefore, semantically, "emergence" here refers to the same definition that is formulated in the study of other - natural or social - complex systems (Goldstein, 1999).

Entrepreneurship is “a complex and multidimensional phenomenon” (Gartner, 1985, p. 696-7), an “organizing process” (Gartner et al., 2010, p 99), in which multiple variables interact. Several scholars have proposed different frameworks to explain the characteristics of the firm creation process (Carter et al., 1996). Gartner’s theoretical framework (1985) for describing new venture creation is particularly suitable for this research that tries to develop a model of venture emergence. According to Gartner, venture creation involves the following aspects (Gartner, 1985):

- Characteristics of the **individual**(s) who start the venture, such as age, education, need for achievement, risk taking propensity, etc. They also include:
    - Locus of control.
    - Job satisfaction.
    - Previous work experience.
    - Entrepreneurial parents or friends or partner.

- The **organization** which they create and its characteristics, organizational structure and strategy of the new venture, such as the new product or service, joint ventures, customer contracts, etc. Other characteristics are:
    - Overall cost leadership.
    - Differentiation.
    - Focus.
    - Parallel competition.
    - Franchise entry.
    - Geographical transfer.
    - Supply shortage.

- Tapping unutilized resources.
- Customer contract.
- Becoming a second source.
- Licensing.
- Market relinquishment.
- Sell off of division.
- Favoured purchasing by government.
- Governmental rule changes.

- The **environment** surrounding the new venture and its conditions and context, such as competitors, venture capital availability, accessibility of suppliers, customers, transportation, etc. Also:
  - Venture capital availability.
  - Technically skilled labour force.
  - Accessibility of customers or new markets.
  - Governmental influences.
  - Proximity of universities.
  - Availability of land or facilities.
  - Accessibility of transportation.
  - Attitude of the area population.
  - Availability of supporting services.
  - Living conditions.
  - High occupational and industrial differentiation.
  - High percentages of recent immigrants in the population.
  - Large industrial base.

- Larger size urban areas.
- Availability of financial resources.
- Barriers to entry.
- Rivalry among existing competitors.
- Pressure from substitute products.
- Bargaining power of buyers.
- Bargaining power of suppliers.

- The **process** by which the new venture is started and the activities undertaken by nascent entrepreneurs during the new venture creation process: location of the business opportunity, accumulation of resources, etc. These are:
  - The entrepreneur locates a business opportunity.
  - The entrepreneur accumulates resources.
  - The entrepreneur markets products and services.
  - The entrepreneur produces the product.
  - The entrepreneur builds an organization.
  - The entrepreneur responds to government and society.

Gartner proposed a list of variables of new venture creation under each different dimension of this framework. The different possible interactions among the variables have the potential of a high degree of complexity and they would explain the “kaleidoscopic” diversity among the processes of the emergence of ventures and the “enormously varying patterns of new venture creation” (Gartner, 1985, p. 701). Gartner’s framework and variables would be extremely useful, especially in the

modelling section, given that it provides a sound way to conceptualize variation and complexity in the context of an agent-based methodology.

A firm is not instantaneously established. The creation of a firm requires performing a series of activities undertaken by nascent entrepreneurs during the organization creation process (Carter et al., 1996). These venture organizing activities “consist of those activities that establish the physical structure and organizational processes of a new firm” (Delmar and Shane, 2003). These activities are performed with great variations, to different degrees, different order, different points in time, and even by different member of the entrepreneurial team. Does the timing these activities determine the survival of new ventures? Although the kinds of activities that nascent entrepreneurs undertake, the number of activities, and the sequence of these activities have an impact on the success of creating a new venture, it is not clear if the sequence itself is significant (Delmar and Shane, 2004).

Empirical research following up Katz and Gartner’s (1988) framework were not able to find a pattern or sequence of events or activities in common to all emerging organization undertaken by nascent entrepreneurs during the organization creation process or organization gestation (Reynolds and Miller, 1992; Carter et al., 1996; Gartner et al., 2010). However, Delmar and Shane (2004) have argued that the timing of undertaking particular organizing activities indeed has an influence in the survival of new ventures. In particular, those nascent entrepreneurs who initially focus their activities on acquiring legitimacy would be in better position for survival. Legitimacy of a new venture is understood as a way in which stakeholders can recognize that the new entity adheres to accepted rules, norms, principles and standards, such as establishing a legal form or writing a business plan (Delmar and Shane, 2004). Legitimacy increases the ability of create social capital, making connections with external stakeholders, establishing external legitimacy through the improvement of

the terms of transactions with other actors (suppliers, clients, investors, etc.), and consolidating internal production procedures for transforming resources (Delmar and Shane, 2004). On the other hand, Shim and Davidsson (2018) found that the higher probabilities of venture emergence are around three months after initiation of the nascent entrepreneurial process. As time goes by, the venture emergence chance decreases and the risk of abandonment and failure increase after seven months.

The most common first activity in the creation of an organization is a personal commitment by nascent entrepreneurs involved in the new venture. The most common last activities in the creation of a new firm were to hire first employees, first sales income, and to get external financial support (Carter et al., 1996). But organizations emerge neither in an orderly periodic progression of activities nor in a random sequence, and that none of the individual gestation activities may be a necessary condition to success (Arenius, Engel and Klyver, 2017). The sequence of start-up activities in venture emergence seems to follow a “chaotic pattern” that points out to a process consisting in a nonlinear dynamical system neither stable or predictable, nor purely stochastic or random (Cheng and van de Ven, 1996). The low-dimensional chaotic pattern of organization emergence suggests a simple nonlinear dynamic systems of only a few variables (Cheng and van de Ven, 1996) that would make possible to develop a meaningful model. Examining the dynamic patterns among these activities using theory and methods from complexity science, Lichtenstein et al. (2007) found that emergence of new firms occurs when the rate of start-up activities is high, they are spread over time, and they are concentrated at a later time in the process of organizing (Lichtenstein, Carter and Gartner, 2007).

The Panel Study of Entrepreneurial Dynamics (PSED) provided a list of organization formation activities, obtained from previous studies on the relationship between nascent entrepreneurial behaviour and the creation of

new firms (Reynolds and Miller, 1992; Gatewood et al., 1995; Carter et al., 1996). These start-up activities in PSED, I and II, ordered by prevalence, were (Reynolds and Curtis, 2008, Table 5.8, p. 214):

- Serious thought given to the start-up.
- Actually invested own money in the start-up.
- Began saving money to invest in the start-up.
- Began development of model, prototype of product, service.
- Began talking to customers.
- Began defining market for product, service.
- Organized start-up team.
- First use of physical space.
- Purchased materials, supplied, inventory, components.
- Initiated business plan.
- Began to collect information on competitors
- Purchased or leased a capital asset.
- Began to promote the good or service.
- Receive income from sales of goods or services.
- Took classes, seminars to prepare for start-up.
- Determined regulatory requirements.
- Open a bank account for the start-up.
- Established phone book or internet listing.
- Developed financial projections.
- Arranged for child care, household help.
- Began to devote full time to the start-up.
- Established supplier credit.
- Legal form of business registered.
- Sought external funding for the start-up.
- Hired an accountant.
- Liability insurance obtained for start-up.
- Established dedicated phone line for the business.
- Initiated patent, copyright, trademark protection.
- Hired a lawyer.

- Hired an employee.
- Received first outside funding.
- Joined a trade association.
- Proprietary technology fully developed.
- Initial positive monthly cash flow.
- Acquired federal Employer Identification Number (EIN).
- Filed initial federal tax return.
- Filed for fictitious name (DBA).
- Paid initial federal social security payment.
- Paid initial state unemployment insurance payment.
- Know that Dun and Bradstreet established listing.

The variables names with the start-up activities and their prevalence (in percentage) has recently been compiled by Reynolds (2017b):

TABLE 1 - TABLE OF START-UP ACTIVITIES PREVALENCE (FROM REYNOLDS, 2017B.):

Table 10. Start-Up Activities Prevalence: U.S. PSED I and II Cohorts

VARIABLE NAME (1)	START-UP ACTIVITY	PSED I	PSED II
	Total Cases	830	1,214
		Percent	**Percent**
THINK_AW	Serious thought	99.9	99.3
ONINVAW1	MBR 1: Invested own money	93.6	80.6

BUSPLNAW	Began business plan	71.2	73.2
MODEL_AW	Developed model, prototype	87.8	75.0
PURCHAAW	Purchased materials, supplies, parts	81.0	70.4
DFNMKTAW	Define markets to enter	92.0	67.9
PROMOTAW	Promote products or services	72.2	61.5
SALES_AW	Sales, income, or revenue	62.7	66.7
LEASE_AW	Leased, acquired major assets	65.5	65.5
TLKCSTAW	Talk to customers	NA	85.4
FINPRJAW	Financial projections	57.0	47.6
FTWK_AW1	MBR 1: Full time start-up work	46.3	29.7
SAVMONAW	MBR 1: Saving money to invest in firm	79.6	NA
PHLISTAW	Phone book listing for business	28.7	63.5
BKACCTAW	Established bank account for firm	54.2	52.7
SUPCRDAW	Obtained supplier credit	49.9	39.1
SUTEAMAW	Began to organize start-up team	66.6	NA
SPACE_AW	First use of physical space	NA	73.1
IFOCPTAW	Collect information on competition	NA	73.1
HIRE__AW	Hire employee	28.4	21.6
FNDREGAW	Determine regulatory requirements	NA	63.8
ASKFNDAW	Asked for formal funding	34.5	28.3

CSHFL_AW	Cash flow covers expenses, not owners	30.5	24.7
FEDTAXAW	Federal income taxes	42.2	49.0
FICA AW	Federal social security payment (U.S.)	29.2	28.6
LEGAL_AW	Legal form registered	NA	48.8
EIN AW	Acquired registration number	NA	36.1
HRACCTAW	Hire accountant	NA	39.8
CLASS_AW	Took class, seminar, workshop	56.3	NA
DBA_ AW	Acquired doing business as name	NA	29.4
PATENTAW	Patent, copyright, trademark filing	26.6	10.1
PHLINEAW	Business phone line established	33.7	NA
TDASOCAW	Joined trade association	NA	22.4
GETFNDAW	Got initial formal financing	12.2	20.0
CLDCARAW	Arranged child care, housekeeping	41.8	NA
LIABISAW	Obtained liability insurance	NA	30.9
UNEMP_AW	Filed state unemployment ins (U.S.)	18.4	14.7
HELPPRAW	Contact with helping program	24.7	NA
HRLAWRAW	Hire lawyer	NA	26.3
BUSPLFIAW	Business plan finished	NA	47.5
PRDCPLAW	Model, prototype fully developed	NA	46.4
EQTAGAW1	MBR 1: Signed ownership agreement	NA	11.9

PRTECHAW	Proprietary technology developed	NA	12.0
ONINVAW2	MBR 2: Invested own money	NA	24.9
FINSPTAW	Investment in legal business	NA	21.4
DANDB_AW	Know listed in Dun & Bradstreet (US)	7.2	8.5
EQTAGAW2	MBR 2: Signed ownership agreement	NA	12.0
FTWK_AW2	MBR 2: Full time start-up work	NA	9.6
ONINVAW3	MBR 3: Invested own money	NA	6.0
GOTPNTAW	Received patent, copyright, trademark	NA	6.0
EQTAGAW3	MBR 3: Signed ownership agreement	NA	5.8
EQTAGAW4	MBR 4: Signed ownership agreement	NA	3.4
ONINVAW4	MBR 4: Invested own money	NA	2.6
FTWK_AW3	MBR 3: Full time start-up work	NA	2.1
EQTAGAW5	MBR 5: Signed ownership agreement	NA	1.3
ONINVAW5	MBR 5: Invested own money	NA	0.9
FTWK_AW4	MBR 4: Full time start-up work	NA	0.5
FTWK_AW5	MBR 5: Full time start-up work	NA	0.0

Source: (1) Based on common names in the consolidated data sets.

(2) NA = not asked in the interview.

Table of start-up activities prevalence (from Reynolds, 2017b.)

2.1.4. The Individual Aspects Of Venture Emergence: Nascent Entrepreneurs' Human and Social Capital and Their Opportunity Recognition.

Eventually, this project will develop an agent-based model to understand the emergence of new ventures. One of the key agents of the model is the concept of **nascent entrepreneur**. This subsection will briefly address three important aspects of this individual agent that would have to be taken into account in the model: human and social capital of the nascent entrepreneur, the concept of opportunity, and the mechanisms of recognition and exploitation of this opportunity.

HUMAN CAPITAL

The nascent entrepreneur brings two types of human capital -- knowledge and skills - to the new venturing project. On one hand, the general human capital, such as age, genetics, personality, overall education and life history and work experience, and, on the other hand, the specific human capital related directly to the tasks involved in organization creation (Dimov, 2010). Davidsson and Honig (2003) showed that general human capital made more probable the engagement in venturing, although it was not a good predictor of business success.

Among the components of nascent entrepreneur's human capital, there are two very specific to the organization creation: experience in previous venture start-up processes, and knowledge and acquaintance of the industry or sector (Dimov 2010). These aspects of the human capital help to the process of venture emergence, and, although the "tacit, procedural knowledge" acquired through prior previous entrepreneurial and industry experience are important resources for the nascent entrepreneur, they did not predict a successful emergence process, but rather an

increase of the frequency of gestation activities over time (Davidsson and Honig, 2003; Dimov, 2010).

SOCIAL CAPITAL

Nascent entrepreneurs also bring their social capital to the new venturing project. Social capital is understood as the beneficial aspects that can be provided by nascent entrepreneur's social structures, networks and memberships such as closed and extended family, community-based or organizational relationships, etc. The effects of social capital have a broad range: it can be provision of concrete resources, like a loan provided by the family, to intangible assets, like the information on a new potential client (Davidsson and Honig, 2003).

Davidsson and Honig (2003), based on the strength of ties, distinguish between "bonding social capital" and "bridging social capital". Bonding social capital is referred to "strong ties", such as having parents or close friends who owned firms, and it increases the possibility of becoming a nascent entrepreneur. Bridging social capital is based on "weak ties", such as being a member of a business network, member of the Chamber of Commerce, Rotary or Lions, etc., and it is a strong predictor of rapid and frequent gestation activities and for carrying the venture emergence further, for example, to a first sale or a profit, and signalling a successful emerging process (Davidsson and Honig, 2003).

OPPORTUNITY RECOGNITION

Together with the concept of "emergence", the idea of "opportunity" has become a fundamental aspect of the phenomenon of entrepreneurship (Shane and Venkataraman, 2000). Nevertheless, there is certain debate

regarding the nature of the process of opportunity discovery and recognition. Some scholars consider that opportunities are real, concrete entities ready to be noticed, discovered and exploited by entrepreneurs. Entrepreneurs' social capital would provide networks that help to discover and exploit opportunities (Davidsson and Honig, 2003). This "opportunity discovery approach" uses an economics framework, giving relevance to alertness and informational asymmetries among individuals. Opportunities, so to speak, come from "outside" of the entrepreneur (Alvarez and Barney, 2007).

On the other hand, other scholars will argue that opportunity should be considered an emergent cognitive and social process - the "social psychological approach" - in which opportunities would depend on entrepreneurs' own abilities, efforts and activities: it would be as a creative process (Gartner et al., 2010). Gartner et al. (2008), using data from PSED I, suggested that the entrepreneurs'' own experience is closer to the "opportunity creation" approach. However, more empirical research is needed (Gartner et al., 2010).

This research would take this debate relative to the nature of opportunities, discovery or creation, from a modelling point of view. The relevant aspect for modelling is that the entrepreneur "encounters" an opportunity, and this encounter in itself is what it counts. It may can from "inside", internal, such as a painter finds an inspirational theme for a canvas, or from "outside", created by exogenous shocks to an industry or a market (Alvarez and Barney, 2007).

2.2. HEAVY-TAILED DISTRIBUTION IN ECONOMICS: ANTECEDENTS

2.2.1. HEAVY-TAILED DISTRIBUTIONS

A power law - also referred as a kind of heavy-tail distributions, Pareto distributions, or Zipf's distributions - is usually expressed as a rank/frequency expression:

EQUATION 1 – POWER LAW

$$F(N) \sim N^{-\alpha}$$

Where F is the frequency of the event, N is the rank (and the variable), and *α*, the exponent, that, in power laws, is constant (In exponential equations, however, the exponent is the variable, such as in $f(x) \sim e^{ax}$) (Newman, 2005; Sornette, 2006; Clauset et al., 2009; Virkar and Clauset, 2014).

Power laws or scaling laws have been observed in several phenomena in economics and finance since its identification by Pareto at the end of the nineteenth century (Gabaix, 2008). A special type of power laws relative to the distribution of the variables is also called a Pareto law - a distributional power law -, where the variable Y expresses the probability of occurrence of event X, and where the exponent **α** is independent of the units in which the law is expressed ($Y = kX^{\alpha}$). The Zipf's law has been defined as a Pareto law with exponent **α** approximately equal to 1.

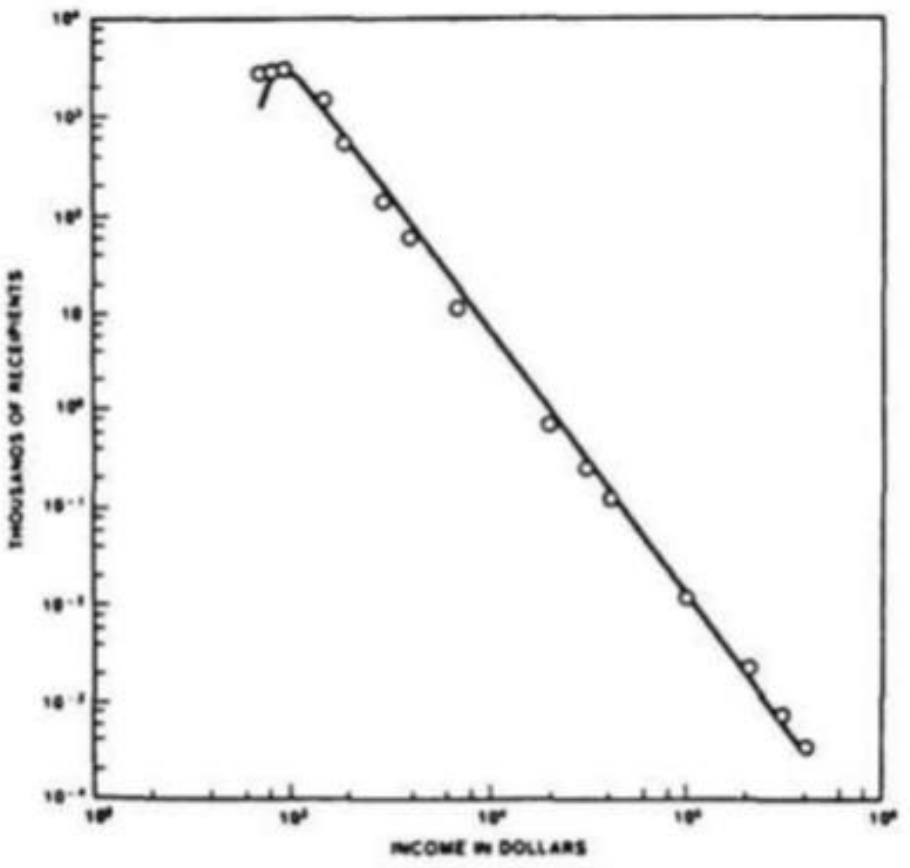


FIGURE 8 - FROM WEST AND DEERING, 1995, P. 173, FIGURE 3.28: FREQUENCY DISTRIBUTION OF INCOMES IN U.S.A. IN 1918.

A typical Pareto's Law distribution of income figure illustrates the distribution of income in a Western country on log-log axes. A straight line with a negative slope signs an inverse power law, with **α** being the slope of the line. Because Pareto found similar power law income distributions for many Western societies, he thought that the slope, α, was a universal constant for western societies, with a value of 1.5, independently, of their particular social structure and institutions. Subsequently, Pareto's assumptions were questioned and several other distributions of income were proposed: Levy, log-normal, Champernowne, Gamma, Boltzmann-Gibbs, and other Pareto variations (West and Deering, 1995; Dragulescu &

Yakovenko, 2001a; 2001b; Brzezinski, 2014; Bee, Riccaboni and Schiavo, 2017). Mandelbrot considered that Pareto's law only applies to the high incomes (Mandelbrot, 1960); Gibrat (1931) proposed that income and wealth distributions were generated by multiplicative random processes, which results in log-normal distributions; Kalecki (1945) insisted that that these log–normal distributions were not stationary, but their width increases in time. Current econophysicists also proposed several versions of multiplicative random processes in order to model and explain theoretically wealth and income distributions (Dragulescu & Yakovenko, 2001a). Eventually, the theoretical justifications of these proposed distributions developed into two schools:

- The socio-economic school: Appealing to economic, political and demographic factors to explains the distribution (for example, Levy, 1987).
- The statistical school: it tries to explain the distributions in terms of stochastic processes, in the econophysics line of research (Dragulescu & Yakovenko, 2001a, 2001b; Tao, 2015; Tao et al., 2017).

2.2.2. Processes For Generating Power Law Distributions

Stumpf and Porter have stated that although one may statistically validated a heavy-tailed distribution, it is necessary to have a theory to support it, i.e. a description of the generative processes that explain it, a model, based on a theoretical framework, that explains the emergence of that distribution (Stumpf and Porter, 2012). Mitzenmacher (2004), Newman (2005), Sornette (2006), and Gabaix (2009) have described several candidate generative processes to explain the emergence of power-law distributions both in natural and social systems, from the simplest algebraic methods to the more complex models, in which critical phenomena are involved.

Although, heuristically, it may be tempting to set aside the more simplest generative processes, more frequent in physics or chemistry, in which randomness - via statistical mechanics - has an important role, this research would also consider them in the second section (agent-based modelling) because, in human social interactions, such as the entrepreneurial nascent processes, chance, luck, randomness, unplanned events, fortuitousness, Gibrat's law events, may play a relevant role (Coad 2009; Coad, 2013; Frankish et al., 2013; Lotti et al., 2009).

Newman (2005) identifies several categories of possible generative models for power laws, starting from the most simple ones, the multiplicative processes - products of random numbers - (Mitzenmacher, 2004) to the more theoretically sophisticated concepts such as Self-Organized Criticality. The capacity of power laws of undergoes several mathematical operations and still give another power law distribution is a remarkable characteristic. For example, feeding an agent-based model with power law distributed inputs, may generate power law distributed outputs, merely for mathematical reasons:

> *"Power laws have very good aggregation properties: taking the sum of two (independent) power law distributions gives another power law distribution. Likewise, multiplying two power laws, taking their max or their min, or a power, etc. gives again a power law distribution. This partly explains the prevalence of power laws: they survive many transformations and the addition of noise." (Gabaix, 2014, p. 9-10)*

> *"One thus expects power laws to emerge naturally for rather unspecific reasons, simply as a by-product of mixing multiple (potentially rather disparate) heavy-tailed distributions." (Stumpf and Porter, 2012, p. 666).*

Combinations of exponentials:

This process has been considered the mechanism to explain the power-law distribution of the frequencies of words, the metaphor of ‘monkeys with typewriters’, or “Monkeys Typing Randomly” (Miller, 1957).

Inverses of quantities:

This mechanism has been used in theoretical physics to explain, for example, magnetic behaviour (Ising model of a magnet) (Sornette, 2006).

Random walks:

In nature, random walks show some properties that are distributed in a power-law form. For example, processes that fluctuate randomly and that end when it hits zero – ‘gambler’s ruin’ - show a power law distribution of the lifetimes. Coad et al. (2013) have applied this process to explain firm growth.

This mechanism has also been used to explain the apparent power law distribution of the lifetime of biological *genera* in the fossil records –and also in other biological taxa – ranks of the Linnaeus’ hierarchy - and branches of the evolutionary trees such as families, orders, and so on.

We will also use partially this mechanism to assign value to some of the variables of the agent-based model of the second section.

The Yule process (or **preferential attachment**):

One of the most applicable processes to understand the presence of power law distributions is the Yule process, developed by G. Udny Yule in the 1920s, in the context of the study of the distribution of the number of species in a genus, family or order, - that also seems to follow power law patterns - (Willis and Yule 1922; Yule, 1925). The Yule process was mathematically improved by Herbert Simon (1955) and it has been used to explain power laws in many different systems such as city sizes (Simon, 1955), paper citations (Price, 1976), links to pages on the internet web (Barabasi and Albert, 1999), city populations, or personal income, becoming the most widely accepted theory for understanding them (Newman 2005).

This type of 'rich-get-richer' process has also be called **Gibrat's rule**, the **Matthew effect** (Merton, 1968), **cumulative advantage** (Price, 1976), or **preferential attachment** (Barabasi and Albert, 1999; Newman, 2005)

Phase transitions and critical phenomena:

This model has been used mostly in physics, addressing what happens to a system when is in the vicinity of continuous phase transitions, also called critical phenomena, critical points, or phase transitions. Percolation transitions, for example, show power law distributions in the mean cluster size in the critical point (Gabaix, 2009; Sornette, 2006).

Self-organized criticality:

> Some dynamical systems are able to arrange themselves to be always at the critical point. These systems self-organize, showing self-organized criticality (Bak, Tang and Wiesenfeld, 1987; Jensen, 1998). Self-organized criticality has been proposed as the generic mechanism to explain the origins of power-law distributions in phenomena such as forest fires (Drossel and Schwabl, 1992), earthquakes (Bak and Tang, 1989), biological evolution (Bak and Sneppen, 1993) avalanches (Bak, Tang and Wiesenfeld, 1987) and other natural phenomena (Bak, 1996; Jensen, 1998).

From a more specific organizational viewpoint, Andriani and McKelvey (2009) also described several additional generative mechanisms (i.e. causal processes) that yield power law distributions such as hierarchical modularity, event bursts, interacting fractals, least effort principle, niche proliferation, etc. (Andriani and McKelvey, 2009). They classified these scale-free theories about causes of power law distributions in four major categories which Crawford et al. reformulated from an entrepreneurship theoretical perspective (2015):

- Positive Feedback mechanisms such as preferential attachment. Given that some firms begin with more resources than others, “Matthew effect” my explain power laws in entrepreneurship.

- Contextual Effects mechanisms such as self-organized criticality (SOC). If a start-up is positioned at a critical point, the addition of a single new input (a patent, an investor) can

cause dramatic change, producing an avalanche of outcomes ("black swan" events, Taleb, 2007).

- Ratio Imbalances mechanisms such as Simon's hierarchical modularity, in which loosely coupled organizations are more adaptable to a dynamic and changeable environment (Simon, 1962).

- Multiple Distributions mechanisms such as in those systems with multiplicative effects, and where the interactions of the parts produce a multiplicative phenomenon rather than an additive one (fractal food webs, positive feedback loops systems, firm and industry size, etc.).

Dealing with complex social phenomena, like the emergence of firms, the most important of these are 1) the Yule process (preferential attachment), 2) the critical phenomena and the associated concept of self-organized criticality (Newman, 2005) and 3) the multiplicative processes. These processes are able to produce power law distributions, and they can also be modelled using agent-based modelling techniques (Epstein, 1999).

Historically, the first Pareto law was referred to income and wealth. Vilfredo Pareto gathered data on wealth and income through different countries and epochs and noticed that the distribution of income and wealth among the population followed a power law: approximately 80% of the wealth was owed by 20% of the population (Pareto, 1896). Schumpeter, commenting Pareto's Law and his contribution to economics, wrote:

"Few if any economists seem to have realized the possibilities that such invariants hold out for the future of our science. (…) In particular, nobody seems to have realized that the hunt for, and the interpretation of, invariants of this type might lay the foundations of an entirely novel type of theory" (Schumpeter, 1949, p. 155-6) (Also in Gabaix, 2008, and Gabaix, 2009).

From the empirical point of view, several power laws have been suggested in economics. The principal mechanism proposed to explain distributional power law in economics – Pareto's distributions - has been proportional random growth (Gabaix, 2009). Proportional random growth generates distributional power laws. Using Yule mathematical theory of evolution (1925), Champernowne (1953) and Simon (1955) applied this mechanism in economics. The work of Champernowne (1953), Simon (1955) and Mandelbrot (1963) explored these distributions in different areas such as firms sizes, cities sizes, and income, and opened a new research path based on stochastic growth that has been followed since then (Sutton, 1997; Luttmer, 2007).

However, the explanation of the stability of the Pareto exponent in different economies, societies and epochs is still under discussion (Gabaix, 2008; Bee, Riccaboni and Schiavo, 2017). Power laws in economics also appear, for example, in city sizes (Gabaix and Ioannides, 2004), salaries of executives (Gabaix and Landier, 2008), in stock market activities such as returns, trading volume and trading frequency (Gopikrishnan et al., 1999; Gopikrishnan et al., 2000), or even in the distribution of macroeconomic disasters worldwide (Barro and Tao, 2011). Many of these empirical regularities, with the current economic theories apparatus, has not been explained yet (Gabaix, 2009; Gabaix 2014). Several attempts have been made to introduce the concepts, methods and models of statistical mechanics, dynamical systems and complexity to address them in the context of this new multidisciplinary branch of economics called "econophysics" (Stanley et al., 2000; Stanley and Plerou, 2001; Durlauf,

2005; Rosser, 2008; Holt, Rosser and Colander, 2011; Buldyrev, et al., 2013).

2.2.3. Firm Size Distributions

Firm size distributions are the results of many complex interactions among several economic forces: entry of new firms, growth rates, business cycles, business environment, public regulations, etc. The underlying dynamics and explanations that drives the distribution of firms' size is still an issue under intense debate (Zambrano, 2015; Bee, Riccaboni and Schiavo, 2017) and organization scholars are discussing which distribution – log-normal, Pareto, Weibull or a mixture of them - is the best-fitting (Gaffeo et al., 2012). The distribution of firms' sizes seems to follow a Zipf's law (i.e. a power law with an exponent close to 1), and this regularity holds for different methods for measuring firm sizes (number of employees, assets, market capitalization) and different countries (In USA: Axtell, 2001; Gabaix and Landier 2008; Luttmer 2007; In Europe: Fujiwara et al., 2004; In Japan: Okuyama et al., 1999). However, there are significant deviations from the Zipf's distribution for the very small and the very large, and for different industrial sectors – the lower and the upper tails of the firm size distribution - (Cabral and Mata 2003; Marsili, 2005; Marsili, 2006; Cefis, Marsili and Schenk, 2009).

Previous research, before the 2000s, although using partial data – only firms listed in the stock market -, was also able to identify Zipf's laws in firm sizes (Ijiri and Simon, 1974; Stanley et al., 1995). These studies were mainly conducted over data sets at a very high of aggregation that included large firms in multiple industrial sectors. For example, Hart and Prais (1956) studied the U.K. manufacturing industry, and Simon and Bonini (1958) and Hall (1987) focused on the U.S. manufacturing firms across all sectors (Bottazzi and Secchi, 2006).

Simon (1955) described a stochastic mechanism that produced a distribution similar to Pareto's law, a model with similar underlying structure of Champernowne's (1953) (Simon 1955). These mechanisms assume that the process satisfies "Gibrat's law": "*all firms have the same expected growth rate and the same standard deviation of growth rate*" (Gabaix, 2014, p. 6).

In the 1930s, the French engineer Gibrat proposed the first formal model of the dynamics of firm size and industry structure to explain the empirically observed size distribution of firms (Sutton, 1997), taking on the following assumptions:

(a) The growth rate of a firm is independent of its size (also known as "the law of proportionate effect").

(b) The successive growth rates of a company are uncorrelated in time.

(c) Firms do not interact (Gibrat, 1931).

It has also been defined a Gibrat's law for means ("the mean of the growth rate is independent of size"), and a Gibrat's law for variance ("the variance of the growth rate is independent of size") (Gabaix, 2009).

Gibrat's firm size distribution regularity was not studied in depth until the 1950s and 60s, when several models were proposed combining Gibrat's Law with other assumptions and caveats (Sutton, 1997). This generation of models based on "stochastic growth" culminated with the works of Simon and his co-authors in the late 70s. Their models modified

Gibrat's assumptions to better fit the empirical data and defined the market as a sequence of independent opportunities, which arise over time.

Simon and Bonini's model (1958) was one of the first attempts of finding an economic explanation to the regularity in the size distribution of firms. Instead of the traditional explanation based on the static cost curve that was not able to predict the distribution of firms by size and has not explanation of the observed Pareto distribution, they proposed a theory based on a stochastic model of the growth process. They assumed the Gibrat's law - the law of proportionate effect - , that is, that size has no effect upon the expected percentage growth of a firm: a firm with assets a billion dollars' worth has the same probability of growing, for example, 20%, as a firm with a million dollars in assets (Simon and Bonini, 1958, p. 609):

> *"It has been shown (Simon 1955) that the Pareto curve can be derived from Gibrat's law, which states that the percentage growth rate of a firm is distributed independently of its size." (Ijiri and Simon, 1974, p. 316)*

Without the assumption of the law of proportionate effect – Gibrat's law, or an approximation to it - distributions from stochastic processes do not generate highly skewed distributions such as the log-normal, the Pareto distribution, the Yule distribution, or others (Simon and Bonini, 1958). The law of proportionate effect is a central feature of Simon and Bonini's model. Successive models trying to explain and model firms' size distribution will retain this concept or adapt it (Ijiri and Simon 1964; Sutton, 1997; Gabaix, 2009).

The second key basic assumption of Simon and Bonini (1958) model – being the first the law of proportionate effect - is that "*new firms are being born in the smallest-size class at a relatively constant rate*" (Simon and Bonini, 1958, p. 610). This assumption of a constant "birth rate" for new

firms determine the generation of Yule/Pareto distributions, instead of log-normal ones. An economic interpretation for the parameter **α** of the power law was proposed: "*it measures, in a certain sense, the rate of new entry into the industry*" (Simon and Bonini, 1958, p. 615). They also called for "*a new statistical measures of the degree of concentration and new interpretations of the economic implications of concentration*" (Ijiri and Simon, 1964, p. 77). Thus, the slope of the Pareto curve should be understood as a measure of the degree of business concentration in an industry or an economy (Cefis, Marsili and Schenk, 2009).

Simon and Bonini (1958) foresaw the potential public policy implications of the processes involved in the firm sizes distribution that, determined by a stochastic dynamics, can be altered through different administrative interventions, and they proposed to re-examine the principles of public policy based on static equilibrium economic schemes developing stochastic models of economic growth instead (Simon and Bonini, 1958; Durlauf, 2012). Degree of industrial concentration – for example, via mergers and acquisitions -, antitrust policies, or monopoly inefficiencies are pertinent examples of major issues related to firm distributions and their growth dynamics (Lucas, 1978; Cefis, Marsili Schenk, 2009).

In 1964, Ijiri and Simon (1964) developed an improvement in the stochastic model for firm sizes distributions, in which they "weakened" one of the key assumptions of the model in order to obtain a more consistent one closer to the observed facts: they introduced some variations into the law of proportionate effect or Gibrat's law. Instead of considering that the probabilities of the size changes are independent of a firm's present size, that is, each firm has the same probability as any other firm of increasing or decreasing in size by any amount in year-to-year changes (say, 5%, 10%, etc.) – a Markoff process -, they reformulated the assumption applying the Gibrat's law only to firm size groups or strata, and not to individual firms (Ijiri and Simon, 1964). They observed that, in reality, the rate of change in size,

are not equal for all individual firms. Year-to year changes in firm size showed different percentage variance, decreasing with increase in size. “Strong” Gibrat’s law compliance only was observed with whole size groups, such as in different industry groups (Ijiri and Simon, 1964). Gibrat’s law, assumed in a weak form, was also able to produce skewed equilibrium distribution (Ijiri and Simon, 1964).

Surprisingly, now that agent-based modelling is a trend in natural and social sciences (Farmer and Foley, 2009), we can already glimpse the concepts and principles of agent-based modelling in the Ijiri and Simon’s simulations (1964), experimenting with the growth patterns produced by the model through additional “runs”, following the individual firms – agents - at successive time intervals and changing the parameters (Ijiri and Simon, 1964). Their results were decisive in the confirmation of the plausibility of stochastic modelling and explanation for the Pareto/Yule distribution of firm size, based on the “*size independence of percentage growth rate (Gibrat’s law) and constancy of the entry rate*” (Ijiri and Simon, 1974, p. 317).

The 1964 model of Ijiri and Simon was still too simple and did not include the effect of mergers and acquisitions, or the possibility of a decrease in size of individual firms (Ijiri and Simon, 1964). Later, it was discovered that mergers and acquisitions indeed do affect the Pareto distribution, increasing the concavity of the curve and introducing a significant departure from the theoretical Pareto distribution (Ijiri and Simon, 1974; Cefis, Marsili and Schenk, 2009). However, when the entire population of firms is considered, mergers and acquisitions do not affect the global Pareto size distribution and remains invariant, that is, Pareto law may only hold when we focus on aggregate statistics (Cefis, Marsili and Schenk, 2009).

This new framework of stochastic growth would be decisive in the later literature, although it was forgotten for almost a decade (Sutton, 1997). With new empirical findings, showing that within an industry, smaller firms grow faster and are more likely to fail than large firms, literature on this tradition had a revival during the mid-80s, such as the Jovanovic's Bayesian learning model (1982) (Bottazzi and Secchi, 2006). In Jovanovic's model, firms learn about their efficiency as they operate within an industry: "*the efficient grow and survive; the inefficient decline and fail*" (Jovanovic, 1982, p. 649).

Nevertheless, there were a certain discontent with the "pure stochastic" character of the models of the 1950s and 1960s. Rather, the aim was to develop standard, conventional maximizing models with the mere introduction of some stochastic elements into them (Sutton, 1997). On the other hand, from an empirical point of view, Gibrat's law was highly controversial and different studies had shown that it may not hold: Gibrat's law contrasts with many theories of firms' growth and it is at odds with other empirical data (Caves 1998; Cefis, Ciccarelli and Orsenigo, 2007). Eventually, Sutton (1997) formulated new empirical facts that do not always were in agreement with Gibrat's assumptions:

1. The probability of survival increases with firm size (Hopenhayn, 1992, p. 1141; Caves 1998, p. 1957).
2. The proportional rate of growth of a firm conditional on survival is decreasing in size (Evans, 1987; Hall, 1987; Cabral and Mata, 2003, p. 1075).
3. For any given size of firm, the proportional rate of growth is smaller according as the firm is older, but its probability of survival is greater (Caves 1998, p. 1959).
4. It is frequently observed that the number of producers tends first to rise to a peak, and later falls to some lower level.

Systematic departures from the Pareto Law can also be observed when the analysis is at the sectorial level, or at specific industrial sectors – in contrast to the aggregate level -. Concavity and different distributional forms appear - such as the log-normal -, and technology play a relevant role in shaping firm size distributions (Dosi et al., 1995; Marsili, 2005). In the Pareto distribution, the size of small firms is underestimated and the size of large firms is overestimated (Marsili, 2006). On the other hand, the distribution seems to change over time (Cabral and Mata, 2003), be affected by recessions, institutional changes and other macro-economic events (Marsili 2006), and may differ from a lognormal-like distribution, evolving over time toward symmetry (Cabral and Mata, 2003).

Some authors have argued that the apparent regularities of the Pareto distribution and the Gibrat's law are simply statistical "artefacts", the results of the aggregation of multiple data, which conceals the high heterogeneity in firm size distribution and the real dynamics of industries across different sectors (Bottazzi and Secchi, 2006). However, further research is needed regarding the evolution of the firm size distribution over time at different levels (global, sectorial, etc.) (Cabral and Mata, 2003; Marsili, 2006).

Other different models have been proposed that have tried to improve the drawbacks of Gibrat's Law assumptions. Bottazzi and Secchi (2006) presented a model that tried to avoid the implicit Gibrat's assumption that firms' growth processes are independent, that there is no form of competition among firms. Although Bottazzi and Secchi (2006) still used the random, stochastic, Simon-inspired tradition on firm dynamics, they built a model in which a stylized idea of competition is introduced: "*luck is the principal factor that finally distinguishes winners from losers among the contenders*" (Bottazzi and Secchi 2006, p. 236). The idea of competition is implemented through the assignment procedure of different business

opportunities among different firms. The probability of obtaining new opportunities depends of the number of opportunities already caught by the firm. In this way, they introduced the “increasing returns” feature in the growth process of firms, a characteristic of the possible diverse positive feedback mechanisms observed within markets, business and industries: economies of scope, economies of scale, networks possibilities, knowledge accumulation, etc. It would be like a version of the “preferential attachment” mechanism applied to business opportunities.

Luttmer (2007) was able to obtain the observed firm size distribution based on entry and fixed cost, firm-specific preference and technology shocks, and selective survivals of firms. Entering firms were able “to imitate” in order to success (Luttmer, 2007). The mechanism used random growth and Brownian motion similarly to the model developed by Gabaix (1999) for the city size distribution (Luttmer, 2007). In this model, the observed Zipf’s law distribution is interpreted “to mean that entry cost are high or that imitation is difficult, or both.” (Luttmer, 2007, p. 1103). The small size of the entrant firms points out that imitation is not an easy task for companies.

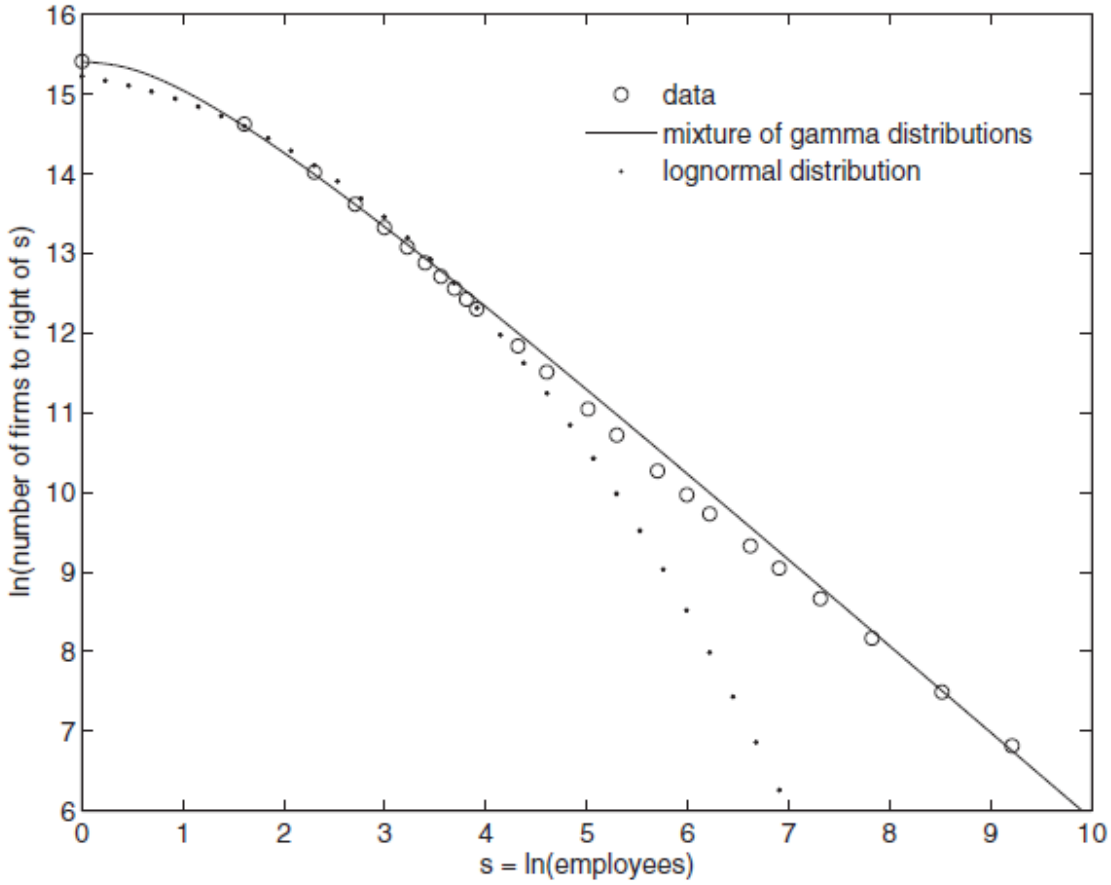


FIGURE I
Size Distribution of U. S. Firms in 2002

FIGURE 9 - FROM: LUTTMER, E. G. (2007). SELECTION, GROWTH, AND THE SIZE DISTRIBUTION OF FIRMS. *THE QUARTERLY JOURNAL OF ECONOMICS*, P. 1104.

In addition to the several proposed models for the random growth of firms such as Luttmer (2007), Gabaix (2009; 2014) has also proposed another mechanism that may also play a relevant role in economics: efficiency maximization. In biology, the energy that an animal of mass M requires to live (metabolic rate) is proportional to $M^{3/4}$. West et al. (1997) proposed that the explanation is related to optimization, to maximize physiological efficiency: the $M^{3/4}$ law emerges because the optimal network system to send nutrients to the animal is a fractal (scale-free) system. Gabaix (2014) has posed the question that if in economics, optimization may also explain the network of power law distributed firms: “does it come from optimality, as opposed to randomness?” (Gabaix 2014, p. 15). Hence,

stochastic processes may not be the main reason to observe power law distribution in firms' sizes or outcomes. Optimal organization may also be a decisive factor, and this property may arise in economics as much as in biology (West, 2017).

Zambrano et al. (2015) - following the econophysics and econochemistry movement in a paper titled "Thermodynamics of firms' growth" - have presented a new thermodynamic model based on the Maximum Entropy Principle that tries to describe the dynamics and distribution of firms' growth. They explain the empirical exponent of Pareto's law as the capacity of the economic system for creating or destroying firms. If the exponent is larger than 1, creation of firms is favoured; when it is smaller than 1, destruction of firms is favoured; if it is equal to 1 (Zipf's law), the system is in a full macroeconomic equilibrium, allowing free creation or destruction of firms. They expect to build a formalized theory based on thermodynamics of the evolution of firms that would lead to a clear and intuitive interpretation of the exponents, and to find a tool for making better diagnosis of the health of an economy and facilitating the development of improved public policies on fair competence and antitrust measures (Zambrano, 2015).

3. HEAVY-TAILED DISTRIBUTIONS IN NASCENT ENTREPRENEURIAL PROCESSES.

3.1. HEAVY-TAILED DISTRIBUTIONS CLASSIFICATION: THE RELEVANCE OF THE PROPER TAXONOMY OF THE ENTREPRENEURIAL EMPIRICAL DISTRIBUTIONS.

The proper classification (taxonomy) of an empirical distribution has an enormous relevance: it reveals the generative mechanism of an organizational process and how and why it emerges (Joo, Aguinis and Bradley, 2017). The accurate identification of a distribution has major implications for appropriately understanding and, eventually, modelling a complex process such as nascent ventures' emergence (Virkar and Clauset, 2014). To get statistical evidence for or against a certain distribution is complicated, especially if we have large fluctuation at the tail of the empirical dataset (Virkar and Clauset, 2014). Therefore, the classification of a dataset is not a straightforward task and it may require the combination of graphical and statistical tests to reach a desired level of confidence in analysing real data (Clauset et al., 2009; Cirillo, 2013). On the other hand, the identification of the best fit distribution in nascent entrepreneurial outcomes is not a trivial matter: it affects the foundations of theory and practice in the research of entrepreneurship (Crawford, Aguinis, Lichtenstein, Davidsson, and McKelvey, 2015). It may be critical for entrepreneurial theory development, testing, modelling, forecast and practice.

As it was mentioned above, non-normal heavy-tail distributions have captured the attention of researchers in different disciplines (West, 2017), and their study have produced important theoretical and practical innovations in several fields, such as physics, computer science,

biomedicine, and economics (Mitzenmacher, 2004; Newman, 2005). For example, in the identification of power law distributions, Mitzenmacher (2005) proposed that the following issues should be addressed:

(a) *Observation*: Collection of the data on the behaviour of the system and demonstration that a heavy tail distribution appears to fit the data sets.

(b) *Interpretation*: Explanation of the significance of the distribution behaviour to the system.

(c) *Modelling:* The proposal of an underlying model that explains the distribution behaviour, for example, with the use of Agent-based Modelling and Simulation (ABMS).

(d) *Validation:* Data validation of the model, including the necessary modifications of the model and its parameters.

(e) *Control:* to control, modify, and improve the system behaviour using the understanding from the model.

In the last years, several statistical methods for fitting heavy-tailed distributions have been developed lowering the barriers for classification that involves complex mathematical procedures, sophisticated algorithms and elaborated code writing (Clauset et al., 2009; Ginsburg, 2012; Alstott et al., 2014; Gillespie, 2015). This paper directly benefits from these new fitting packages, especially those from Joo, Aguinis and Bradley (2017) and Gillespie (2015), both based in Clauset's et al. methods (Clauset et al., 2009). Otherwise the analysis of the different international data sets would have been extremely complicated and time consuming (Limpert and Stahel, 2011).

The goodness of fit of a distribution requires comparing it with the fit of other distributions; in this case, using log-likelihood ratios to identify

which of the several potential fits are better (Joo, Aguinis & Bradley, 2017). Methodologically, it is not necessary to know if a distribution exactly follows a certain function or not, but rather if the distribution considered is the best description available of the real data set. Systems in the real world have noise, and, therefore, few empirical processes should be expected to follow a theoretical mathematical distribution (Alstott et al., 2014). On the other hand, observed data come from a specific real system and the generative mechanism of that system produced the observed data. The candidate distribution and its associated generative mechanism have to be plausible to the system and the processes that we are analysing. If the candidate distribution does not offer a meaningful and credible generative mechanism, there is no reason to use it to describe a real data set.

When studying non-normal heavy-tailed distributions in real entrepreneurial data sets, initially, most of the methodological approaches have assumed the pure power law distribution as the main hypothesis (Crawford et al., 2014, 2015). However, not all non-normal heavy-tailed distribution fit a pure power law (Aguinis et al., 2016). Not until very recently, software improvements have increased the precision of the analysis, providing new data treatment procedures that makes much easier to explore the better fits for a given heavy-tailed distribution.

The objective of this research – and future research - is to analyse the entrepreneurial outcomes of several panel studies on nascent entrepreneurs in different countries in order to discover if they follow any distinct distribution, taking into consideration the different types and families of non-normal heavy-tailed distribution. Previous research has focused mainly in US and Australia data (US PSED and Australian CAUSEE): this paper will introduce also the analysis of other countries – those with their panel datasets are in the public domain - in order to explore if a worldwide pattern may exist.

The theoretical framework that this research uses is the distribution taxonomy developed by Joo, Aguinis & Bradley (2017), and it will be applied to nascent entrepreneurial outcomes. This research will also consider how these distributions are associated with an idiosyncratic generative mechanism, and it will explain how the identified generative mechanism may work in the entrepreneurial processes. Methodologically, we will use the distribution pitting techniques newly developed in R (software package "**Dpit**") also by Joo, Aguinis & Bradley (2017) (freely available on http://www.hermanaguinis.com or on the Comprehensive R Archive Network – CRAN -) which is able to compare many distributions types and to assess how well each distribution may fit a given data set.

Our study, extended across different panel study in different countries, suggests that lognormal is the more common distributions in entrepreneurial outcomes. In some datasets, the power law with an exponential cutoff distribution seems a better fit, but the *p* value makes very difficult to discern if the difference is statistically significant with regard to the lognormal distribution fit (see table 2 in the Appendix).

If lognormal distributions are pervasive in nascent entrepreneurial outcome, generative mechanisms that are not consistent with this distribution are not meaningful for explaining the process. Former research in entrepreneurship pointed out to the prevalence of pure power law distribution in the outcome variable and its generative mechanism (for example, self-organized criticality).

3.1.1. Classification/Types of Non-Normal Heavy-Tailed Distribution in Entrepreneurial Outcomes

Joo, Aguinis and Bradley (2017) have proposed a new taxonomy of distributions in organizational literature, consisting in seven possible total distributions, grouped into four general categories:

(1) Pure power law.

(2) Lognormal.

(3) Exponential tail (including exponential and power law with an exponential cutoff).

(4) Symmetric or potentially symmetric, including Normal, Poisson, and Weibull (Clauset, Shalizi and Newman, 2009).

Most of natural - and even social - phenomena can be described by these seven functions (Limpert, Stahel and Abbt, 2001; Sornette, 2006). And each distribution category can usually be explained by a particular generative mechanism: pure power law by self-organized criticality (Bak, 1996), log-normal distributions by proportional differentiation (Limpert et al., 2001), exponential tail distributions by incremental differentiation (Amaral et al., 2000; Nirei & Souma, 2007), and symmetric distributions (Normal, Poisson, and Weibull) by homogenization processes. These four generative mechanisms are mutually exclusive and, therefore, they may contribute greatly to a better development of theory and modelling (Joo, Aguinis & Bradley, 2017).

TABLE 2 - TAXONOMY OF JOO, AGUINIS & BRADLEY (2017) WITH THEIR GENERATIVE MECHANISMS

General Distribution Category	Generative Mechanism
pure power law distributions	self-organized criticality
log-normal distributions	proportional differentiation
exponential tail distributions (including exponential and power law with an exponential cutoff)	incremental differentiation
symmetric distributions (Normal, Poisson, and Weibull)	homogenization processes

.

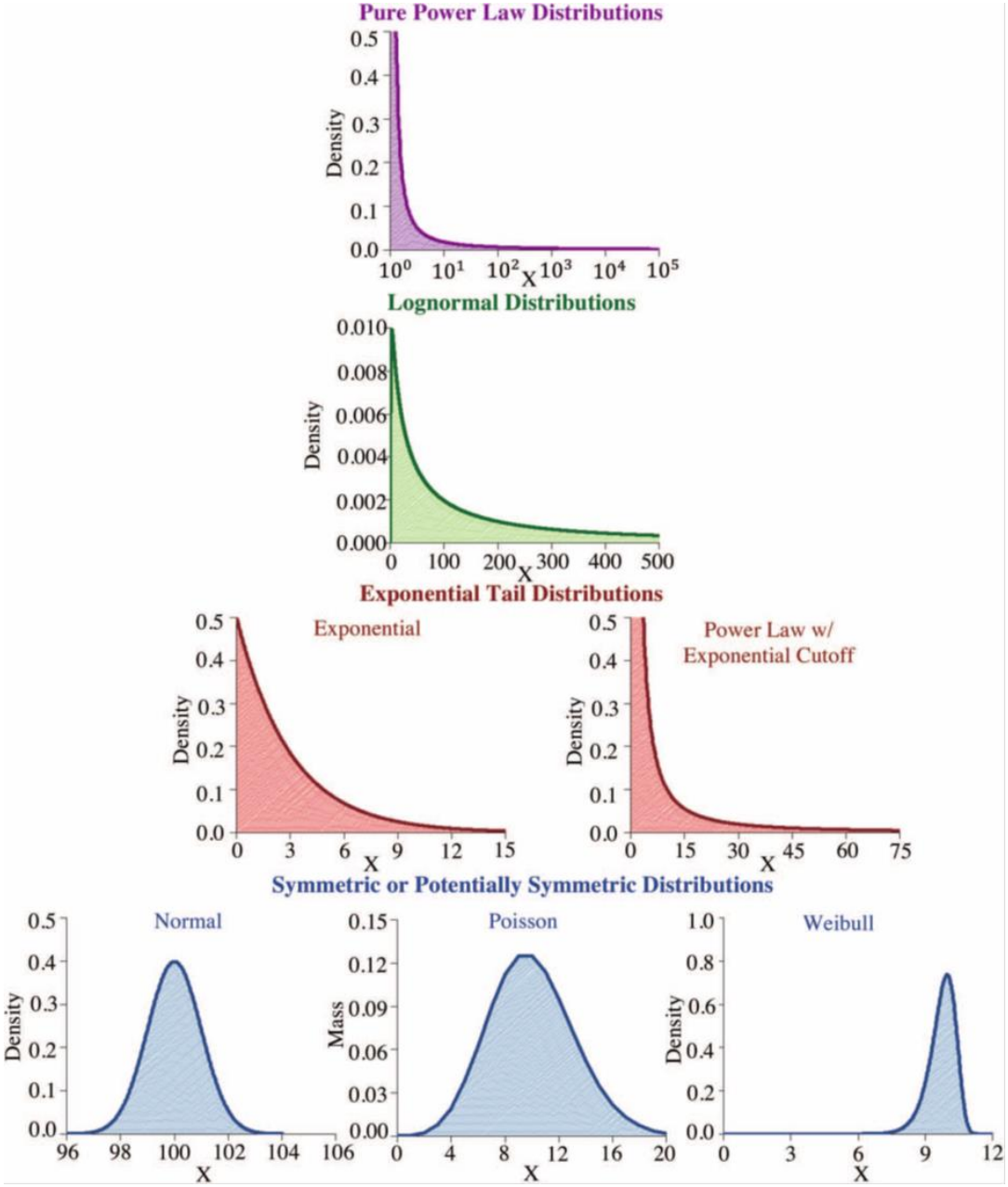


FIGURE 10 - VISUAL REPRESENTATION OF THE TAXONOMY OF OF JOO, AGUINIS & BRADLEY (2017) WITH SEVEN MAIN TYPES OF DISTRIBUTIONS.

Parameters: Pure power law (α = 1.5); log-normal (μ = 5, σ = 2); exponential (λ = 0.5); power law with an exponential cut-off (α = 1.5, λ = 0.01); Normal or Gaussian (μ = 100, σ = 1); Poisson (μ = 10); and Weibull (β = 20, λ = 10). The x-axis represents values of a continuous variable and the y-axis is the probability of a given value or range of values, except for the Poisson distribution, in which the x-axis is a discrete variable, and the y-axis the probability of the discrete variable taking on a given discrete value.

Although Joo, Aguinis & Bradley (2017) introduced seven distributions, the next section will show that the empirical study on nascent entrepreneurial outcome variables data sets across different countries only detected two main prevalent distributions: log-normal, and until certain extent, power law with exponential cut-off (or Weibull, although it is difficult to confirm it). This research will follow the “theory pruning” approach, focusing only in those processes that are able to generate the more pervasive distributions in our nascent entrepreneurial data sets (Leavitt et al., 2010). Then, we will explore the theoretical and practical implications for the entrepreneurial process of these distributions and their generative mechanisms, applying the methodology and data processing described by Joo, Aguinis & Bradley (2017).

4. MATERIALS AND METHODS

4.1. INTERNATIONAL LONGITUDINAL PANEL STUDIES AND VARIABLES

Although there are several theoretical and conceptual difficulties to measure size and growth of nascent ventures and their performance, given the variables currently available in the empirical data sets, this research will follow Crawford and McKelvey (2012) study, taking as the main variables revenues and the number of employees (Cooper, 1993; Coad 2009). Specifically, this study analysed these two outcomes variables - revenues and number of employees - in different nascent entrepreneurial panel studies across three countries located in three different continents: USA, Australia and Sweden. Only these four data sets (Australia, Sweden, and U.S. PSED I & II), out of the 14 projects that have already been implemented, are currently publicly available (Reynolds, 2017b).

4.1.1. PANEL STUDIES OF ENTREPRENEURIAL DYNAMICS II (PSED II) – USA.

As described above, PSED II (started in 2005) was an improved replication of PSED I, and it makes a series of follow-up interviews of an initial cohort of 1,214 American nascent entrepreneurs (Reynolds and Curtin, 2008). These longitudinal studies were also replicated in other countries - such as Australia, Canada, China, Latvia, Netherlands, Norway, UK and Sweden - over the past decade, and they still are at different stages of development (Reynolds and Curtin, 2011).

Full details on all interview schedules and questionnaires of the Panel Study of Entrepreneurial Dynamics, as well as codebooks and

complete data sets are freely available on the project website at the University of Michigan: http://www.psed.isr.umich.edu

From the PSED II, this research considers ten variables related to new venture outcomes, measured by the number of employees and annual revenues at every yearly wave, wave B to F:

These variables are - as defined in the code book by Curtin (2012) -:

Total Revenues:

- PSED II USA - Total Revenues BV2
- PSED II USA - Total Revenues CV2
- PSED II USA - Total Revenues DV2
- PSED II USA - Total Revenues EV2
- PSED II USA - Total Revenues FV2

Number of regular Employees:

- PSED II USA – Number of regular Employees BU2
- PSED II USA - Number of regular Employees CU2
- PSED II USA - Number of regular Employees DU2
- PSED II USA - Number of regular Employees EU2
- PSED II USA - Number of regular Employees FU2

4.1.2. The Comprehensive Australian Study of Entrepreneurial Emergence Research Project (CAUSEE)

Inspired by the American PSED, the Australian CAUSEE follows a sample of approximately 600 emerging start-ups, firms that are in the process of being established (nascent firms), and another sample of approximately 600 newly established young firms (Davidsson and Steffens, 2011). The four annual waves of data collection were completed in 2007/8 - 2010/11 (Davidsson, Steffens and Gordon, 2011). There is extensive documentation on the dataset in the related codebook (Gruenhagen et al., 2016). The CAUSSE datasets, documentation and reports are freely available at:

https://www.qut.edu.au/research-all/research-projects/the-comprehensive-australian-study-of-entrepreneurial-emergence-causee

https://eprints.qut.edu.au/49327/

The variables studied in this research were:

CAUSEE Australia Number of full-time Employees Young Firms – Wave 1 (Year 1) Variable Name: W1: Q205#
CAUSEE Australia Number of full-time Employees Young Firms – Wave 2 (Year 2) Variable Name: W2_B16

CAUSEE Australia Number of full-time Employees Young Firms – Wave 3 (Year 3) Variable Name: W3_B16
CAUSEE Australia Number of full-time Employees Young Firms – Wave 4 (Year 4) Variable Name: W4_B16
CAUSEE Australia Number of full-time Employees Young and Nascent Firms – Wave 5 (Year 5) Variable Name: W5_Q24
CAUSEE Australia Number of full-time Employees Nascent Firms – Wave 1 (Year 1) Variable Name: W1: Q252#
CAUSEE Australia Number of full-time Employees Nascent Firms – Wave 2 (Year 2) Variable Name: W2_C79
CAUSEE Australia Number of full-time Employees Nascent Firms – Wave 3 (Year 3) Variable Name: W3_C79
CAUSEE Australia Number of full-time Employees

Nascent Firms – Wave 4 (Year 4) Variable Name: W4_C79
CAUSEE Australia Sales in $ (Total) (Last 12 Months) Young Firms – Wave 1 (Year 1) Variable Name: W1 Q2027#
CAUSEE Australia Sales in $ (Total) (Last 12 Months) Young Firms – Wave 2 (Year 2) Variable Name: W2_B18
CAUSEE Australia Sales in $ (Total) (Last 12 Months) Young Firms – Wave 3 (Year 3) Variable Name: W3_B18
CAUSEE Australia Sales in $ (Total) (Last 12 Months) Young Firms – Wave 4 (Year 4) Variable Name: W4_B18
CAUSEE Australia Sales in $ (Total) (Last 12 Months) Young Firms – Wave 5 (Year 5) Variable Name: W5_Q18 [&R32] [note: same as NF]
CAUSEE Australia Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 1 (Year 1) Variable Name: W1 Q2030#

CAUSEE Australia Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 2 (Year 2) Variable Name: W2_C85_consolidated
CAUSEE Australia Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 3 (Year 3) Variable Name: W3_C85
CAUSEE Australia Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 4 (Year 4) Variable Name: W4_C85_consolidated
CAUSEE Australia Sales in $ (Total) (Last 12 Months) Nascent and Young Firms – Wave 5 (Year 5) Variable Name: W5_Q18[& R32] Misma variable que YF

4.1.3. Swedish Panel Study of Entrepreneurial Dynamics (Swedish PSED).

Similarly to the US PSED II and the Australian CAUSEE, the Swedish PSED followed 623 nascent entrepreneurs during a six-year period (Samuelsson, 2011; Honig and Samuelsson, 2012). The data sets are freely available in Dr. Samuelsson's ResearchGate page.

https://www.researchgate.net/project/Swedish-PSED

This page includes:

- Samuelsson, Mikael. Dataset: erc-neo-ne6-n12-n18-n24-proj— project based data file will all waves from month 0 to month 14, SPSS.SAV file available on Research Gate.
- Samuelsson, Mikael, Dataset: ERC/PSED-75. 75 month follow up data. Technical Report: SWE PSED codebook— all variables with names and waves.

Also:

Delmar, Frederic. Data form the Swedish PSED (n=223), 1998-2000.

https://www.researchgate.net/publication/266630741_Swedish_PSED_Final_Data_1998

Delmar, Frederic. Coding Manual for file Swedish PSED data 1998.

https://www.researchgate.net/publication/262796537_Coding_manual_for_Swedish_PSED_final_data

The variables studied in this research were:

SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave 1 (Year 0) Variable Name: gw31nn00
SWEDISH PSED Number of full-time Employees– SWE PSED 1 Wave 2 (6 months) Variable Name: gw31nn06
SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave 3 (12 months) Variable Name: gw31nn12
SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave 4 (18 months) Variable Name: gw31nn18
SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave 5 (24 months) Variable Name: gw31nn24
SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave N75 (75 months) Variable Name: gw31n

SWEDISH PSED - Outcome Variables Sales Turnover (Thousands SEK) Last Year Variable Name: pt11nn18
SWEDISH PSED Sales Turnover (Thousands SEK) First 3 Months Variable Name: pt12nn18
SWEDISH PSED Sales Turnover (Thousands SEK) First 6 Months Variable Name: pt13nn18
SWEDISH PSED Sales Turnover (Thousands SEK) First 12 Months Variable Name: pt14nn18
SWEDISH PSED Sales Turnover (Thousands SEK) Second year of operation (24 months) Variable Name: pt11nn24 (global dataset)
SWEDISH PSED Sales Turnover (Thousands SEK) Sales Turnover in 1997 Variable Name: pt31nn24 (global dataset)
SWEDISH PSED Sales Turnover (Thousands SEK)

Sales Turnover in 1998 Variable Name: pt21nn24 (global dataset)
SWEDISH PSED Sales Turnover (Thousands SEK) Last Year Sales Turnover after 75 months. Variable Name: pt11n (N75 SPSS file)
SWEDISH PSED 35. Sales Turnover (Thousands SEK) Second year of operation (24 months) file SPSS erc-n24 Variable Name: SWE_pt11nn24_erc-n24
36. Sales Turnover (Thousands SEK) Sales Turnover in 1998 Variable Name: pt21nn24_erc-n24 – ver otro file SPSS erc-n24 SWE_pt21nn24_erc-n24
SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave 5 (24 months) Variable Name: gw31nn24 – Specific dataset SPSS erc-n24
SWEDISH PSED Sales Turnover (Thousands SEK) Sales Turnover in 1997 Variable Name: pt31nn24_erc-n24 – Specific dataset SPSS erc-n24

4.1.4. Direct Access to the Datasets

- ***Australia:***

  Comprehensive Australian Study of Entrepreneurial Emergence (CAUSEE).

- ***Sweden (SE-PSED):***

  https://www.researchgate.net/project/Swedish-PSED

  - Delmar, Frederic. Coding Manual for file Swedish PSED data 1998. Author provided.
  - Samuelsson, Mikael. Dataset: erc-neo-ne6-n12-n18-n24-proj—project based data file will all waves from month 0 to month 14, SPSS.SAV file available on Research Gate.
  - Samuelsson, Mikael, Dataset: ERC/PSED-75. 75 month follow up data. Author provided. Samuelsson, Mikael. Technical Report: SWE PSED codebook—all variables with names and waves, available on Research Gate.

- ***U. S. PSED I, II:***

  All interview schedules, data sets, and codebooks available online.

- ***Five Cohort Harmonized Data Set:***

  Reynolds, P. D., Hechavarria, D., Tian, L.-R., Samuelsson, M., & Davidsson, P. (2016). Panel Study of Entrepreneurial Dynamics: A Five Cohort Outcomes Harmonized Data Set. Research Gate.

4.2. DATA ANALYSIS

The accurate identification of the distribution patterns is complicated because of the large fluctuations in the empirical tail of the data distribution, that make very difficult the distinction from alternative heavy-tailed distributions (for example, the power law, the log-normal or the stretched exponential or the Weibull) (Virkar and Clauset, 2014). The proper identification of a heavy tailed distribution has theoretical implications and it should be statistically validated (Stumpf and Porter, 2012). For example, to

test for the fit of a power law, Clauset et al. (2009) propose a Kolgomorov-Smirnov test, while Gabaix proposes a simpler alternative method (Gabaix and Ibragimov, 2009; Gabaix, 2009).

To calculate the power law model fit and to validate it statically, Crawford and McKelvey (2012) and Crawford et al. (2014; 2015) used the mathematical procedures described by Clauset et al. (2009). Virkar and Clauset (2014) have also developed and updated the protocols for analysing heavy-tailed distributions in binned empirical data. Clauset has developed a companion web page, hosted by the Santa Fe Institute, in which the MATLAB and R scripts (also with other software scripts) of the protocols and techniques are freely accessible at:

http://tuvalu.santafe.edu/~aaronc/powerlaws/
http://tuvalu.santafe.edu/~aaronc/powerlaws/bins/

In the first section of this research, we used and followed procedures and layouts of the R statistical package **Dpit ()** designed by Joo, Aguinis & Bradley (2017) in order to identify the better fit of the heavy tailed distributions of the empirical – real world - datasets. The **Dpit** package aims to identify the better-fit options. However, if we need to study a specific distribution more in detail, we need to use complementary statistical packages, already developed and tested in R. For this task, we will use the R package '**goft'** version 1.3.4 ("Tests of Fit for some Probability Distributions") and the more sophisticated package '**fitdistrplus'** version 1.0-9 ("Help to Fit of a Parametric Distribution to Non-Censored or Censored Data"). Both packages are freely available at the CRAN package repository (https://cran.r-project.org/).

THE DPIT() PACKAGE (IN R)

Distribution fitting is not a straightforward task, especially dealing with non-normal and heavy tailed functions (Stumpf & Porter, 2012). It is necessary to compare distributions with one another in order to find the distribution that better fits a sample. Until very recently, with the statistical software packages available, the implementation of this comparative process among distributions was especially difficult and time-consuming. Lately, Joo, Aguinis, & Bradley (2017) have developed a methodology for distribution fitting and a new R package, called **Dpit**, that is able to compare simultaneously the seven types of distributions of their proposed taxonomy: 1) pure power law, 2) lognormal, 3) exponential, 4) power law with an exponential cutoff, 5) Normal, 6) Poisson, and 7) Weibull. Researchers can then examine the fit of the seven distributions per sample.

The package is freely available on: http://www.hermanaguinis.com and on the Comprehensive R Archive Network (CRAN), with other related packages such as **poweRlaw** (Gillespie, 2015). **Dpit** was also built using the code available at http://tuvalu.santafe.edu/~aaronc/powerlaws/ (Clauset, Shalizi, and Newman, 2009; Virkar and Clauset, 2014). In particular, they mainly borrowed their package from Shalizi's code. However, the Dpit package differs from previous packages in relevant aspects. The Dpit package has several remarkable features that make the comparison procedure among alternative distribution much easier.

Firstly, it has loop functions that automatically clean samples, removing missing cases and zeros. Secondly, it skips over unsuccessful calculations and continues processing the rest of the data. These features were not available in previous packages (for example, **poweRlaw** package in R; Gillespie, 2015). Thirdly, **Dpit** sets the minimum value in a sample to the lowest positive number, and, consequently, try to assess the fit of the distribution not only in the tail, but also in the complete data set. That is, the

Dpit() function does not reject data points that fall below a certain threshold, or x_{min}. This is because the goal of **Dpit**() is to determine whether the complete data set itself follows a certain type of distribution, not whether the tail end (only a fragment) of the data set follows a certain type of distribution. This feature is decisive, because previous methodologies and packages focused only on the tail of the distribution, and, as a result, the data set was incomplete and truncated, and many data points that fell below a certain threshold were rejected. The resulting analysis was then biased and distorted, making impossible to identify clearly the generative mechanism of the distribution.

Complementary Statistical Software Packages

Additionally, to the **Dpit()** distribution comparison package, this research also uses two complementary statistical software packages in R: **Goft** and '**fitdistrplus'**. These packages do not compare distributions, but rather they analyse more in detail an empirical dataset and offer illustrative statistical metrics on a specific distribution (p-value, several goodness-of-fit tests, etc.).

The R **Goft** package, developed by Elizabeth Gonzalez-Estrada and Jose A. Villasenor-Alva, is a straightforward package that provides quick tests for the goodness-of-fit and p-values for the different distribution possibilities: gamma, inverse Gaussian, log-normal, 'Weibull', 'Frechet', Gumbel, normal, multivariate normal, Cauchy, Laplace or double exponential, exponential and generalized Pareto distributions. (https://CRAN.R-project.org/package=goft) (Villasenor and Gonzalez-Estrada, 2009; 2015). This package offer a good first approximation to the best fit looking at the p-value of the test.

The second package, '**fitdistrplus'**, developed by Delignette-Muller and Dutang, (2014, 2015), offer a more complete set of statistics, plots and comparisons among the potential distributions (https://CRAN.R-project.org/package=fitdistrplus). It provides functions to help the fit of a parametric distribution to non-censored or censored data. It also provides maximum likelihood estimation (MLE), moment matching (MME), quantile matching (QME) and maximum goodness-of-fit estimation (MGE) methods (when is possible to performance these calculations).

.

In the second section of this research, in the data analysis of the agent-based model, another useful package is **RNetLogo** (Thiele, 2014) can be used. It provides an interface to the agent-based modelling platform NetLogo. The interface allows to use and access Wilensky's NetLogo (Wilensky 1999) from R (R Core Team, 2018) using either headless (no GUI) or interactive GUI mode. It offers functions to load models, execute commands, and get values from reporters making much easier to transfer big amount of data from the agent-based model, generated by the agent-based platform Netlogo, to the statistical software R. Once the agent-based platform has transferred the dataset to R, it can be processed and analysed with the R software packages decribed above.

Software	Authors	Description	Location
R statistical package **Dpit ()**	Joo, Aguinis & Bradley (2017)	Identification of the better fit of heavy tailed distributions in empirical datasets.	http://www.hermanaguinis.com
R package '**goft'** version 1.3.4	Villasenor and Gonzalez -Estrada (2009, 2015)	"Tests of Fit for some Probability Distributions"	https://CRAN.R-project.org/package=goft
package '**fitdistrplus'** version 1.0-9	Delignette-Muller and Dutang, (2014, 2015)	"Help to Fit of a Parametric Distribution to Non-Censored or Censored Data".	https://CRAN.R-project.org/package=fitdistrplus
RNetLogo	Thiele, 2014.	It allows to use and access Wilensky's NetLogo (Wilensky 1999) from R (R Core Team, 2018)	http://cran.r-project.org/web/packages/RNetLogo/index.html

poweRlaw package in R	Gillespie, 2015.	Statistical software package to analyse heavy-tailed distributions.	CRAN package repository: https://cran.r-project.org/

DPIT() PROCEDURE

Although **Dpit()** has embed an internal function that cleans the sample, in any case, to avoid any potential problem, for this research, data sets were previous cleaned removing zeros (the logarithm of zero is undefined), and codes related to "Do not know" or "NA" (Alstott et al., 2014).

After loading the **Dpit** package and the **PoweRlaw** package, the 48 samples of entrepreneurial outcomes were introduced in R, and then, we entered the command line in R: out <- **Dpit**(data set) for each sample. This command led to comparing all seven distributions with each other per sample (i.e., 21 instances of distribution pitting per sample). See Appendix Table 1: Distribution Pitting Statistics (**Dpit()** Results).

For each comparison between two distributions, the **Dpit** package offers two types of statistics for the data set: a log-likelihood ratio (LR) and its associated *p* value. The log-likelihood ratio (LR) measures the degree to which the first distribution fits better than the second distribution. **Dpit** treats one distribution as the first distribution and the other as the second distributions. A positive log-likelihood ratio means that the first distribution is

a better fit. A negative log-likelihood ratio means that the second distribution is a better fit.

A log-likelihood ratio value of zero establishes the null hypothesis (both distributions in the comparison fit similarly). The *p* value of each log-likelihood ratio reflects the extent to which the presence of a nonzero log-likelihood ratio value can be explained merely by random fluctuations (Clauset et al., 2009). Therefore, in this statistical package, the higher the *p* value, the more probable that the log-likelihood ratio value is simply originated by randomness. Joo, Aguinis and Bradley (2017) and Clauset et al. (2009), adopted the *p* value cut-off of 0.10, and considered *p* values higher that 0.10 not statistically significant. When comparing among the different potential distributions, if only one type of distribution was never the worse fit (log-likelihood ratio and p-value), it was considered the probable dominant distribution for that concrete the nascent entrepreneurial data set.

Joo, Aguinis and Bradley (2017) developed a sophisticated protocol to decide which would be the distribution that better fit in the cases that the distribution pitting results were not conclusive. Their data sets were very diverse - 229 samples - and they were able to identify many different types of distributions (such as Weibull, Normal, exponential, etc.). However, unlike the Joo, Aguinis and Bradley (2017) data sets, the nascent entrepreneurial datasets of this study largely showed only two possible dominant distributions: lognormal and power law with exponential cut-off. In many cases, the *p* high values (randomness, noise) made it inconclusive to identify the best fit between lognormal and power law with exponential cut-off distributions.

4.3. RESULTS

Table 1 (see Appendix 1: “Distribution Pitting Statistics (Dpit() Results”) shows the complete detailed distribution pitting statistics generated by the package **Dpit**. There each log-likelihood ratio value and its p-value (p-value in parentheses) can be found for the 21 comparison in total between potential distributions, for each of the nascent entrepreneurial dataset variables.

Table 3 (below) show an example of the **Dpit** results (distribution pitting) for one of the outcome variable of the CAUSEE panel study in Australia, Sales in $ (AUD), in the first wave (first year of the study) for nascent ventures (firms in the process of establishing) (sample number 16). The software compares each of the seven potential distributions with each other. The abbreviation *NormvPL* means the comparison between the normal distribution versus the pure power law distribution (“Norm v PL”). A positive result of the normalized log-likelihood ratio value implies that the first distribution indicates a superior fit in the comparison abbreviation name “*NormvPL*”. On the other hand, a negative result of the normalized log-likelihood ratio value implies that the second distribution (pure power law, *PL*) is the superior fit.

For example, the results corresponding to the comparison between the power law with exponential cut-off **versus** the lognormal distribution (abbreviated as “CutvLogN”) were -3.61 (0.0003). That means that value of the normalized log-likelihood ratio is -3.61 and the p-value is 0.0003. The log-likelihood ratio value is negative; therefore, the second distribution of this comparison “CutvLogN”, the log-normal distribution, should be preferred. The p-value is 0.0003, is below the 0.10-cutoff, implying that the lognormal distribution (LogN) is a better fit in comparison to the power law with exponentical cut-off (Cut) and that the comparison is statistically significant and not only due to randomness in the sample. If we analyse the

rest of the 20 comparisons for this CAUSEE outcome variable (Sales in $ (AUD), in the first wave, nascent firms), with their log-likelihood ratios and their *p* values, we can conclude that, in this specific sample, the lognormal distribution is indeed the best fit, and this result is statistically significant.

TABLE 3 – DPIT () RESULTS FOLLOWING PROCEDURE AND DATA LAYOUT OF JOO, AGUINIS AND BRADLEY (2017) :

Table 3: Dpit results - with procedure and data layout of Joo, Aguinis and Bradley (2017) - (distribution pitting) for one of the outcome variable of the CAUSEE panel study in Australia, Sales in $ (AUD), in the first wave (after the first year of the study) for nascent ventures (firms in the process of establishing) (sample number 16).

The six columns of the table show the comparison results calculated by the software package **Dpit**() in R. For each comparison, it is shown the normalized log-likelihood ratio value followed by the normalized p-value (in parentheses).

Abbreviations of distribution names: PL = Pure power law, LogN = Lognormal, Exp = Exponential, Cut = Power law with an exponential cutoff, Norm = Normal, Pois = Poisson, and Weib = Weibull.

Abreviations of comparison between distributions: For example, **NormvPL** means Normal distribution **versus** power law distribution. A positive result of the normalized log-likelihood ratio value implies that the first distribution indicates a superior fit in the comparison abbreviation name **NormvPL**. On the other hand, a negative result of the normalized log-likelihood ratio value implies that the second distribution is the superior fit.

p = statistical significance for the normalized log-likelihood ratio value.

Poisson's log-likelihood ratio and p-values are not available for continuous data.

Variable	***N (size of sample)***	**NormvPL**	**NormvCut** **PLvCut**	**NormvWeib** **PLvWeib** **CutvWeib**	**NormvLogN** **PLvLogN** **CutvLogN** **WiebvLogN**	**NormvExp** **PLvExp** **CutvExp** **WeibvExp** **LogNvExp**	**NormvPois** **PLvPois** **CutvPois** **WeibvPois** **LogNvPois** **ExpvPois**
16. Sales in $(AUD) (Total) (Last 12 Months) **Variable Name: W1 Q2030#** **Nascent Firms – Wave 1 (Year 1)**	**302**	**-6.59 (0)**	**-8.94 (0)** **-138.8 (0)**	**-7.24 (0)** **-3.70 (0.0002)** **6.08 (0)**	**-8.57 (0)** **-14.27 (0)** **-3.61 (0.0003)** **-11.09 (0)**	**-12.59 (0)** **1.82 (0.07)** **4.47 (0)** **2.56 (0.01)** **4.55 (0)**	**2.41 (0.016)** **2.41 (0.016)** **2.41 (0.016)** **2.41 (0.016)** **2.41 (0.016)** **2.41 (0.016)**

Table 2 of the Appendix (“Distribution Pitting Conclusions”) shows the probable dominant distributions for the 47 (one sample is duplicated) selected nascent entrepreneurial outcome data sets from Australia, Sweden and USA, and some comments about whether the pitting results were statistically significant or not, and their probable generative mechanisms. In many cases, the log-likelihood ratio value pointed out to certain distributions but the p-values are too high, which cast doubts about the rejection of the null hypothesis, that is, about the conclusion that one distribution is indeed a better fit than the other.

The analysis of the distribution pitting results in entrepreneurial outcome datasets showed that both the lognormal and the power law with an exponential cut-off were identified as the best fitting distribution for most of the samples. However, in the comparison between these two distributions, again, sometimes, the p-values were high and, therefore, we could not reject the null hypotheses, that is, that both distributions would have a plausible similar fitting: both may be acceptable statistically. In order to resolve these ambiguous situations, in which, because of the high p-value we cannot determine the best fit - in this case, between lognormal or the power law with a exponential cut-off -, Joo, Aguinis and Bradley (2017) developed a set of decision rules.

The first decision rule of Joo, Aguinis and Bradley (2017) has already been described above: taken into account the positive or negative value of the normalized log-likelihood ratio (being LR=0 the null hypothesis), and the p-value (being the p value cut-off of 0.10), per variable, if only one type of distribution – say, the lognormal - was never the worse fitting distribution, then, that distribution should be considered the probable dominant distribution. However, if several types of distributions were never identified as being the worse fitting option, then, Joo, Aguinis and Bradley (2017) proposed to apply two additional decisions rules.

Their second rule, based in the principle of parsimony, discriminated among nested distributions. Their taxonomy has three pairs of nested distributions. They are:

(a) power law with an exponential cut-off (two parameters) and pure power law distribution (one parameter);

(b) power law with an exponential cut-off (two parameters) and exponential distribution (one parameter);

(c) Weibull distribution (two parameters) and exponential distribution (one parameter).

The second rule states that the distribution with more parameters in the nested distribution is the worst fitting option for the observed dataset. However, in our example, given that the lognormal distribution and the power law with an exponential cut-off are not a pair of nested distribution, we cannot apply this second rule in order to decide which distribution is better.

The third decision rule is also based in the principle of parsimony, and it refers to the distribution with fewer possible “shapes”. In Joo, Aguinis and Bradley’s taxonomy (2017), this is related to their classification in terms of the flexibility of the distribution, that is, the possibility to change the shape of the distribution (for example, the skewedness) merely changing the value of their parameters:

-flexible distributions: the lognormal, Poisson, and Weibull distributions.

- Inflexible distributions: the pure power law, exponential, power law with an exponential cut-off, and normal distributions.

The third decision rule states that when we have to decide between a flexible and an inflexible distribution, we should consider the inflexible distribution the best explanation: it is better to choose the distribution with fewer possible distribution shapes.

These decision rules, however, have strong methodological limitations:

> *"(…), the three decision rules that we used for implementing distribution pitting should not be interpreted as leading to clear-cut, black-and-white results. Instead, our decision rules are designed to help the user choose the most likely dominant distribution for a given dataset, given that the shape of an individual output distribution may be the result of multiple mechanisms operating simultaneously. In the future, methodological advances may allow the user to identify and weigh the importance of each mechanism contributing to the shape of an individual output distribution." (Joo, Aguinis and Bradley, 2017, p. 1043).*

In our samples, the Joo, Aguinis and Bradley's (2017) third decision rule would prefer the power law with an exponential cut-off because it is inflexible, and they would recommend to reject the lognormal distribution because it is flexible. Overall, our results show a high level of inconclusiveness about the best fit distribution, even applying the three rules described by Joo, Aguinis and Bradley (2017). The third rule seems to impose a very strong restriction that forces us to choose only a distribution – power law with an exponential cut-off - in ambiguous situations without a clear methodological background to do so, or forcing us to choose a generative mechanism – incremental differentiation - that is always not coherent with the rest of results. If we consider only rule #1 and #2, the lognormal distribution would be plausible for most of the distributions (except a few variables in the Australian CAUSEE data set). On the other hand, it is possible that a sample may have a section of the data behaving in a log-normal manner, and another section showing a power law

distribution, and that the software is still not developed enough to differentiate both sections in the same sample.

4.4. DISCUSSION OF THE RESULTS

Joo, Aguinis and Bradley (2017) have proposed a new distribution pitting methodology for the assessment of the types of non-normal distributions that are better in the fitting of individual output distributions (Joo, Aguinis and Bradley, 2017). We have followed their methodology for nascent entrepreneurial outcomes datasets across different longitudinal studies in different countries. The implementation of the distribution pitting was through a new R statistical package, called **Dpit**. They also developed a set of decision rules to identify the more dominant function (or functions) and generative mechanism in each sample. After applying the **Dpit** package to the outcomes variables of nascent entrepreneurial datasets, we found that the results mostly suggested two types of distributions for these entrepreneurial samples: power law with an exponential cut-off and lognormal distributions.

However, the results were not completely conclusive. Deciding between a lognormal distribution and a power law with an exponential cut-off distribution, Rule #3 suggests choosing the less flexible distribution, which is the power law with an exponential cut-off. However, at level of Rule #2, choosing lognormal behaviour would offer a more inclusive and plausible common mechanism for explaining the complete set of samples. Choosing the power law with an exponential cut-off would leave behind almost half of the variables, many of which are definitively lognormal, and with a very strong statistical significance.

Except four samples in the Australian CAUSEE panel, which undoubtedly are power law with an exponential cut-off distributions, the complete set of entrepreneurial outcomes variables can be plausibly explained by lognormal distributions and its generative mechanism, proportionate differentiation. Our results in entrepreneurial outcome variables contrast with those of Joo, Aguinis and Bradley's (2017) results in individual outputs in other organizational contexts. They found that 75% of the samples in different occupations and collectives suggested the exponential and power law with an exponential cut-off, and their associated generative mechanism - incremental differentiation - as the prevailing distribution and explanation, that is, some individuals have bigger linear increases in output than others.

The fact that nascent entrepreneurial outcome variables seems to follow lognormal distributions has relevant theoretical and practical implications, and it may reveal that a different generative mechanism than in other organizational processes might be at play. If we may have to reject pure power law distributions and their generative mechanism (self-organized criticality) for explaining the entrepreneurial dynamics, at least in most of the nascent entrepreneurial dataset, how can we explain the emergence of new ventures based on proportionate differentiation, the generative mechanism of lognormal distributions?

5. THEORETICAL AND PRACTICAL IMPLICATIONS: LOGNORMAL DISTRIBUTIONS VERSUS EXPONENTIAL DISTRIBUTIONS

5.1. LOGNORMAL DISTRIBUTIONS

5.1.1. DESCRIPTION OF THE LOGNORMAL DISTRIBUTION

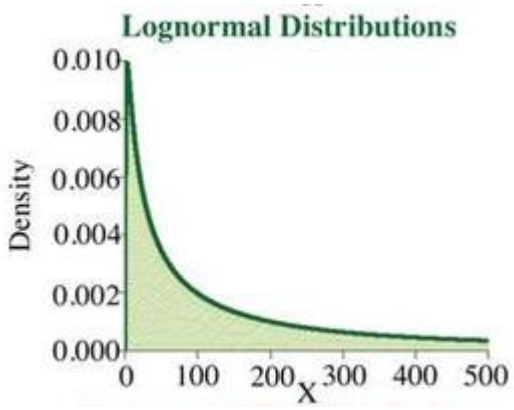


FIGURE 11 - FIGURE OF AN EXAMPLE OF LOGNORMAL DISTRIBUTION FROM JOO, AGUINIS & BRADLEY (2017, P. 1024).

[μ = 5, σ = 2]

The lognormal distribution belongs to the class of the skewed distributions. Many measurements in natural and social sciences show skewed behaviour, especially when mean values are low, variances are large, and the variable values cannot be negative (Limpert et al., 2001). Frequently, those skewed distributions are well described by the lognormal distribution pattern (Aitchison and Brown, 1957; Limpert et al., 2001; Antoniou et al., 2004). Lognormal distributions can be found across sciences, such as geology, epidemiology, environmental sciences,

microbiology, linguistic, economics or finance. Income distributions are classical examples in social sciences and economics (Aitchison and Brown 1957; Limpert et al., 2001). In a classical text on log-normality, Aitchison and Brown (1957), for example, claimed that national income across countries shows lognormal distributions characteristics.

The difference of variability between the normal (Gaussian) and lognormal distribution is based on the way in which different forces act independently of one another in a particular process, whereby the effects are **additive** in normal distributions but **multiplicative** in lognormal distributions. That is, the product of many independent positive random variables – equally distributed - produces lognormal distributions, similarly to the central limit theorem, but in its multiplicative version. This is called in probability “the multiplicative central limit theorem” (Limpert et al 2001, p. 344). If the sum of several independent Gaussian variables produces a Gaussian random variable, the multiplication of several independent lognormal variables generates a lognormal distribution.

The properties of lognormal distributions were defined since the XIX century (Galton, 1879; McAlister, 1879; Gibrat, 1931; Gaddum, 1945). If a random variable X is log-normally distributed, then its logarithm (Y= log(X)) has a normal distribution. The variable has to have positive values (log of 0 is not defined) and the distribution is skewed to the left (see figure below from Limpert et al., 2001, p. 344). The lognormal distribution shows a bell shape head on the left and a finite heavy tail on the right (figure 11 a). Using a logarithm scale on a lognormal distribution will generate the well-known bell shape of the normal distribution (see figure 11 b)

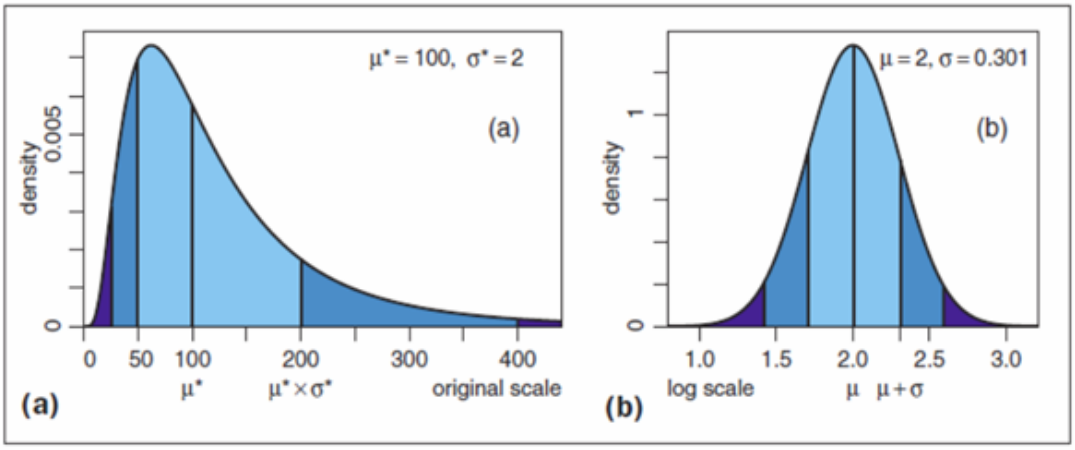


FIGURE 12 - FIGURE 3 FROM LIMPERT ET AL 2001, P. 344. AN EXAMPLE OF A LOGNORMAL DISTRIBUTION WITH ORIGINAL SCALE (A) AND WITH LOGARITHMIC SCALE (B).

Lognormal distributions are Gaussian distributions in the logarithm form of a given variable. If Y is the variable, the distribution appears Gaussian when it is formulated in terms of the logarithm of Y. Similarly to the log Y, the function of the distribution itself in its logarithmic form of the variable is scale invariant. Lognormal distributions arise from non-linear transformations (West and Deering, 1995).

As the normal (Gaussian) distribution, the lognormal distribution is also specified by two parameters of log (Y) of the variable: the mean mu (μ), always positive, and the standard deviation, or sigma (σ) (>0). The mean does not affect the heaviness of the distribution at the right tail; however, a high value of the standard deviation will make the right tail heavier. The ranges of the standard variation are meaningful because they are related to the sources of variability in the processes under study. For example, the lognormal distributions related to the infection processes of pathogens in humans may show different standard deviation depending on the genetic variability of the human populations (Limpert et al., 2001).

To generate a log-normal distribution using a standard statistical software, we have to consider the following steps: If Y represents the variable that we want to have a normal distribution, and μ is the mean and σ the standard deviation of Y, then a log-normal distribution can be code programmed as $e^{\wedge(\text{random-normal M S})}$ where:

$$M = \ln(\mu) - (\beta/2),$$

$$S = \sqrt{\beta}, \text{ and}$$

$$\beta = \ln [1 + (\sigma^2 / \mu^2)],$$

and “random normal” is a procedure that reports a normally distributed random floating point number (Railsback & Grimm, 2012). We will see the relevance of these algebraic expressions to understand lognormal distribution below in the agent-based model coding section of this research.

However, lognormal distributions are difficult to identify (Limpert et al., 2001). The similarities with normal distributions in some aspects are probably the cause of been taken as normal until very recently (Limpert et al., 2001, p. 350). Only because of the new advances in computer software development, their identification has been possible and accessible to natural and social researchers (see above section of new pitting software). Curiously, a normal distribution can be fitted in terms of lognormality with high levels of statistically significance (p-values), but not the opposite. There are not examples of original measurements that follow normal distributions that cannot be described by a lognormal distribution just merely adapting accordingly the parameters (Limpert et al., 2001, p. 350).

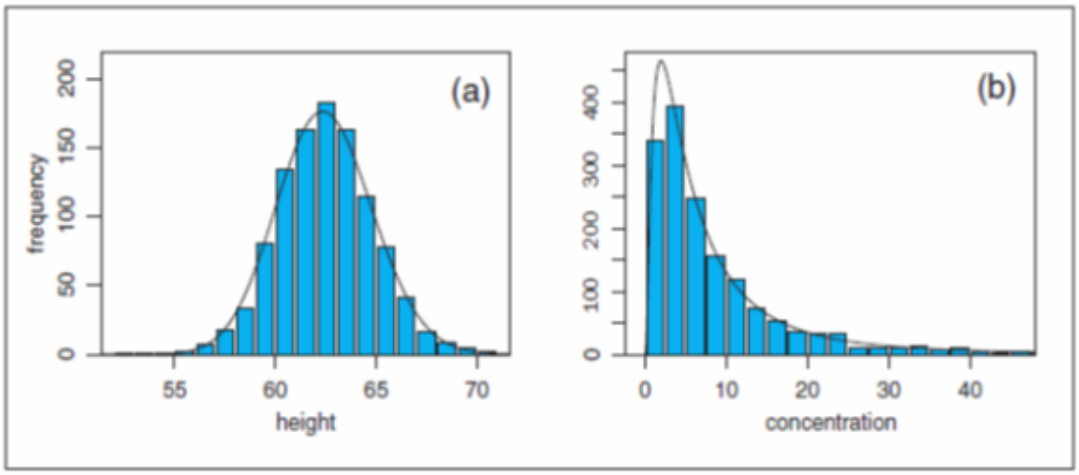


FIGURE 13 - FIGURE FROM LIMPERT ET AL., 2001, P. 342: AN EXAMPLE OF NORMAL DISTRIBUTION (A) AND OF A LOGNORMAL DISTRIBUTION (B). IN THE FIGURE A, THE NORMAL DISTRIBUTION HAS A GOODNESS OF FIT P VALUE OF 0.75, BUT THE LOGNORMAL DISTRIBUTION MAY ALSO FIT EQUALLY WELL WITH A P VALUE OF 0.74. IN CONTRAST, IN FIGURE B, THE LOGNORMAL DISTRIBUTION FIT WITH A P VALUE OF 0.41, BUT NOT WITH THE NORMAL (P VALUE 0.0000).

Normal and lognormal distributions can both describe well a certain dataset when there are small coefficients of variation. However, when we find high level of variability, often, the lognormal distribution is more appropriate.

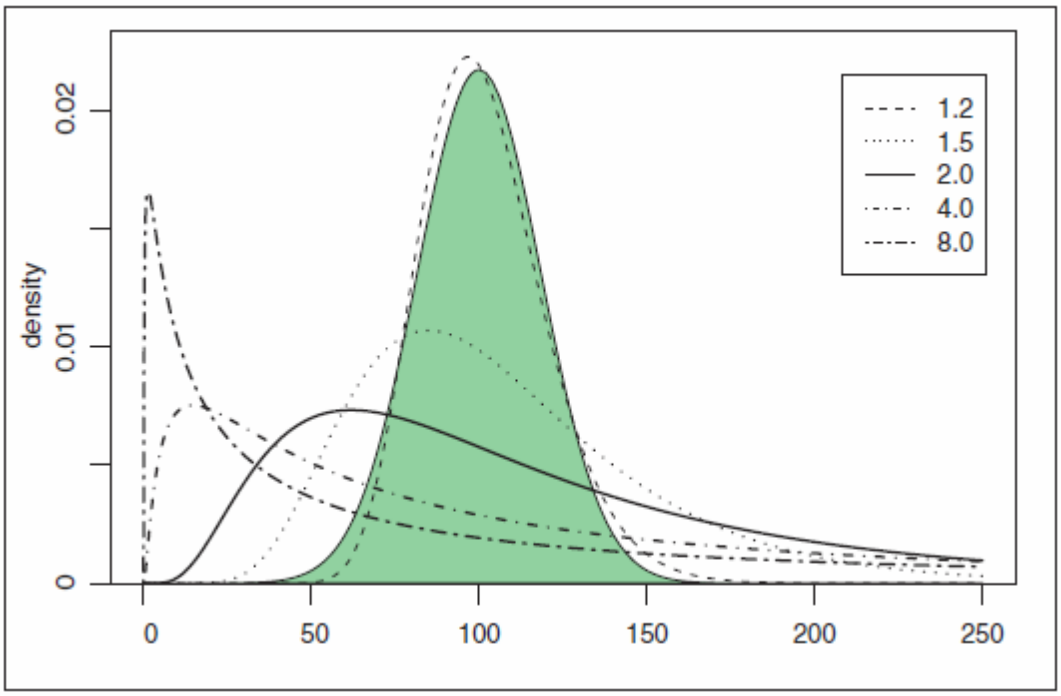


FIGURE 14 - FIGURE 4 LIMPERT, 2011, P. 344. THE FIGURE SHOWS DENSITY FUNCTIONS OF DIFFERENT LOGNORMAL DISTRIBUTIONS COMPARED WITH A NORMAL DISTRIBUTION (SHADED, MEAN = 100; STANDARD DEVIATION = 20). ALL THE LOGNORMAL DISTRIBUTIONS HAVE THE SAME MEDIAN. MERELY CHANGING THE STANDARD DEVIATION OF THE LOGNORMAL DISTRIBUTION, THE NORMAL DISTRIBUTION CAN BE MIMICKED. IT IS POSSIBLE TO GET A NORMAL DISTRIBUTION OUT OF A LOGNORMAL DISTRIBUTION, BUT NOT THE OPPOSITE.

Both pure power laws and lognormal distributions have heavy tails and extreme values on the right, however, lognormal distributions decay more intensely. In comparison with the rest of the distribution, after the pure power law distribution, the lognormal shows the second heaviest right tail.

5.1.2. Lognormal Distributions and Multiplicative Processes

It was mentioned above that lognormal distribution differs from the normal Gaussian distribution because the different forces acting independently on one process are **multiplicative** instead of **additive**. The product of many independent positive random variables – equally distributed - produces lognormal distributions. It is called "*the multiplicative central limit theorem*" (Limpert et al., 2001, p. 344).

But why and how does a lognormal distribution emerge? Aitchison and Brown (1957) described mathematically several theories that may explain the genesis of lognormal distributions. They insist that a distribution with a good fit regarding the empirical data is not always enough. The search for the fundamental base of a distribution may provide clearer insights of the underlying process and it may offer a wider application of the system under study. On the other hand, it helps us to be able to know, understand and modify the distribution parameters in order to meet new circumstances and different empirical data of a similar process elsewhere (Aitchison and Brown, 1957).

Based on the works of Kapteyn (1903) and his analogue machine to generate lognormal histograms, Aitchison and Brown (1957) developed a formulation of "the law of proportionate effect", proposed as the generative mechanism of these distributions:

> "*A variate subject to a process of change is said to obey the law of proportionate effect if the change in the variate at any step of the process is a random proportion of the previous value of the variate*" (Aitchison y Brown 1957, p. 22).

It should be noticed that although the law has been usually considered as an ordered sequence of events in time, especially in the context of biological research – for example, during the period of growth to maturity of an organ or organism -, in other fields such as economics, this approach may be misleading. One variation of the law of proportionate effect states that the greater the number of steps in the sequence, that is, the longer the law of proportionate effect is in operation, the greater the value of the variance (σ^2 parameter). This approach assumes that the law operates continually, *ad infinitum*. But many phenomena, for example, in the study of the size distribution of incomes, this "continuum" assumption cannot be accepted. If the law operates continually, the implication is that the inequalities of incomes, that it is measured by the parameter *variance*, σ^2, must continually increase, which is not the empirical case: the inequality of incomes remains constant through time. To resolve this problem Kalecki (1945) proposed to abandon the assumptions related to consider the processes only in temporal terms. For example, the variations in the inequality of incomes may be considered as mainly determined by multiple economic forces. At any point in time, the distribution of the variable emerges out of an enormous number of causes which operate simultaneously. The outcome of these many different effects and causes, if they interact following the law of proportionate effect, is again to produce a lognormal distribution of incomes (Aitchison and Brown, 1957, p. 25).

As another possible process that may generate lognormal distributions, Aitchison and Brown (1957, p.26-27) introduced the *theory of breakage*, originated in the study of particle-size statistics, based in the works of Kolmogoroff (1941) (later also Epstein, 1947, and Herdan, 1953). Kolmogoroff (1941) proposed this model to explain the emergence of two-parameter lognormal distributions in ores that have been crushed either by natural process or by artificial ones. The theoretical background of

Komogorov's discussion is essentially an application and restatement of the "theory of proportionate effect", mentioned above.

Applying the Russian mathematician Kolmogorov's "theory of breakage", West and Deering (1995) illustrated with an example how to explain the distribution of income. For West and Deering (1995, p. 152), the distribution of income should be understood as a "Multiplicative Statistical Process" that operates as follows. First, it is assumed that to in order to reach a certain level of income, a complex process, several sub-tasks have to be implemented, such as:

1) To be born in a certain social background.
2) To have a minimum educational level.
3) To possess a determined personality type.
4) To be able to perform certain technical skills.
5) To have a certain level of communication skills.
6) To be motivated.
7) To be in the right place at the right time.
8) To be willing of taking risks.

For each individual, a series of probabilities is assigned for the implementation of each of the eight factors above. Thus, the probability of reaching a certain level of total income is proportional to the product of each of these eight probabilities: $p_1 * p_2 * p_3 *\ldots* p_8$. Following Kolmogorov theory, the results will be a lognormal distribution (West & Deering, 1995, p. 152).

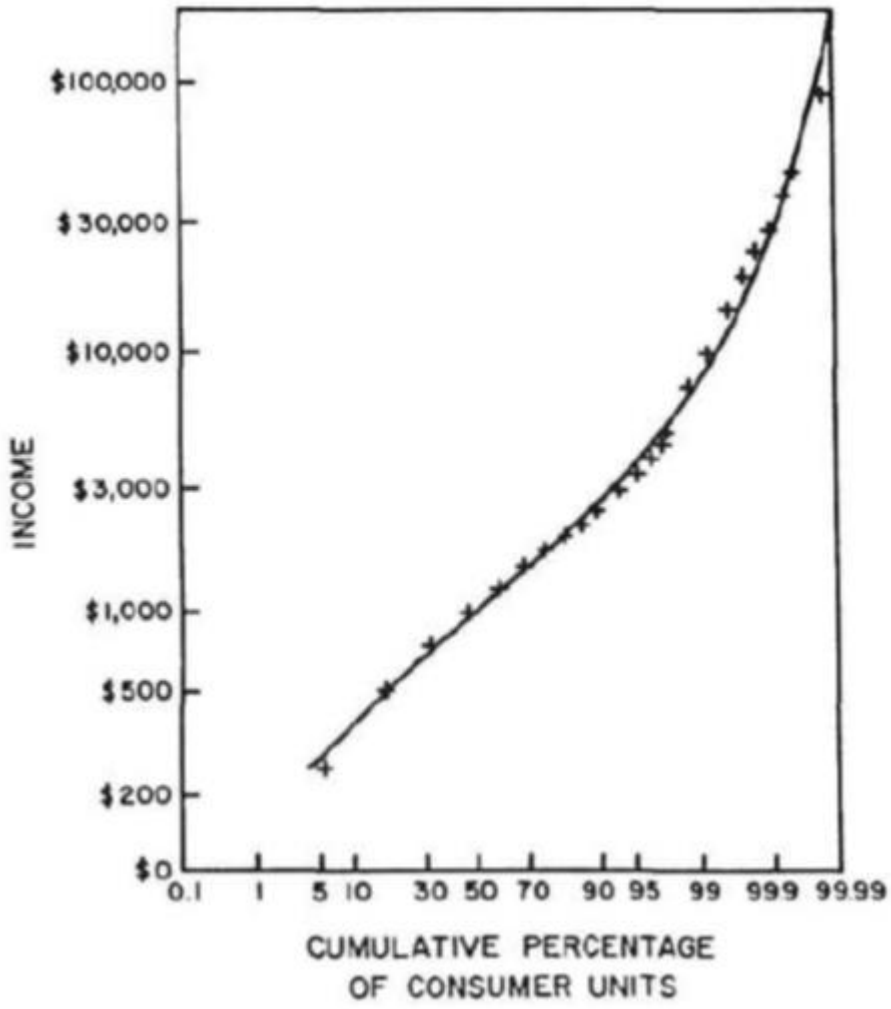


FIGURE 15 - FIGURE PROPOSED BY WEST & DEERING (1995, P. 151) TO SHOW A LOGNORMAL DISTRIBUTIONS OF INCOME LEVELS FOR FAMILIES AND SINGLE INDIVIDUALS IN 1935-36 (FIGURE 3.16)

In this figure (3.16), the lognormal distribution seems to fit adequately for a 97% or 98% of the total population. However, the last 2% or 3% seems to follow an inverse power law (West and Deering 1995, p. 152).

Another example of a multiplicative process generating a lognormal distribution, also introduced by West & Deering (1995), was developed by the controversial Nobel Prize William Shockley in a completely different

context. In the 1950s, Shockley was puzzled for the fact that there was a great difference among the number of scientific publications published by the staff of scientific research laboratories. Some scientists were able to publish at a rate even higher than fifty times more than others were. Shockley also noticed that differences in the rates of performing other human activities are not so big among individuals, for example walk speed (in a range of 2 to 5 mph), running speed, pulse rate, talk speed, etc. show much narrower limits (Shockley, 1957, p. 284). So, why then were there such individual variations of productivity in Research Laboratories that can reach even nearly one hundred-fold between extreme individuals? Why are the spread in rates so greater than it is for other human activities? Shockey (1957, p. 280) argued that in many natural phenomena, where the variables change due to additive effects of a huge number of independently varying factors, a Gaussian – normal - distribution should be expected. However, rates of publication show the normal distribution not in the variable itself, but rather in the logarithm of this rate of publication. Shockley believed that the explanation of these large variations - this statistical peculiarity- is determined by some idiosyncratic characteristics of the creative scientific process and that the lognormal distribution seems a consequence of the way in which the research activities are conducted in a large, modern laboratory (Shockley, 1957). Hence, he proposed the lognormal distribution to explain the complex creative process for scientific research papers and he also developed some models to explain it. The main feature of his models was that a large number of factors are involved in publishing, so very small changes in each of these factors, may result in a very large variation in the creative output. In order to publish a scientific paper in a given period of time, a series of factors and abilities are needed and they require implementation, for example:

1) Ability to consider a good problem or question.
2) Ability to work on this problem or question.
3) Ability to recognize an interesting result.

4) Ability to know and to make the decision of stopping and write up the results.

5) Ability to write properly.

6) Ability to learn from other's criticism.

7) Determination to submit to a journal.

8) Willingness to answer the peer reviewers' objections.

Thus, similarly to the example of the income distribution described above, in this example, each of these paper publishing factor (F_n) have an associate probability, and the total productivity is defined by the product of the probabilities of each factor. The probability that a scientist publishes a paper in a given period of time will approximately be the product of the associate probabilities of the mentioned set of factors, F_1, F_2, etc. related to his/her personal attributes. The total, final publishing productivity of this scientist would then be given by a formula such as

$$P = F_1 * F_2 * F_3 * F_4 * F_5 * F_6 * F_7 * F_8$$

If this model is correct, a small variation in one of these publishing factors (F_n) can produce a large variation in the total publishing productivity of a researcher (Shockley, 1957, p. 286).

To prove his theory, Shockley used data from the staff at the Brookhaven National Laboratory and he could demonstrate that indeed the rate of publication of research papers in physics seems to follow a lognormal distribution.

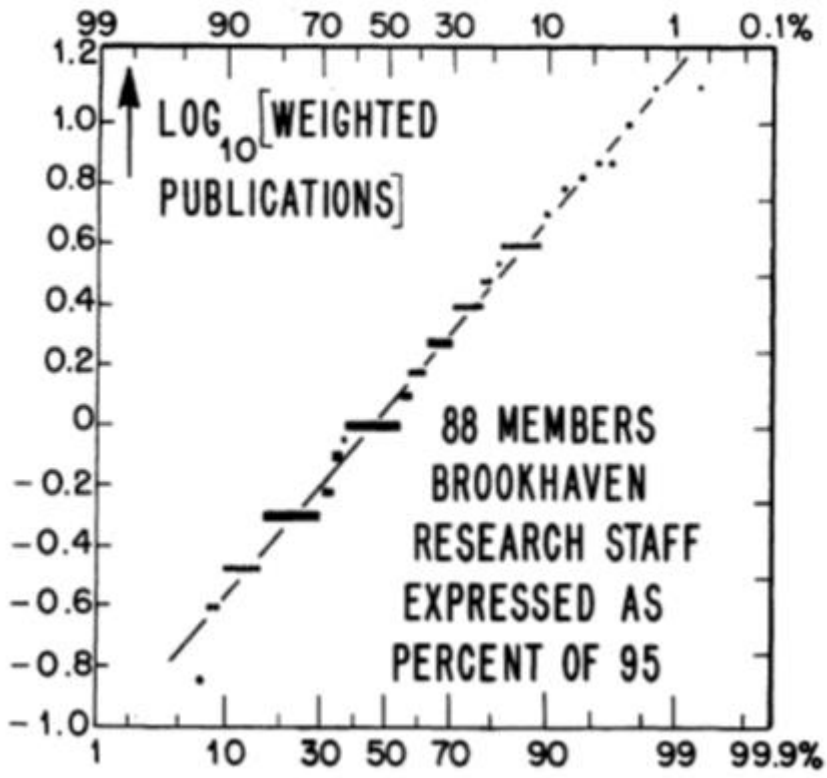


FIGURE 16 - FIG. 7- FROM SHOCKLEY, 1957, P. 283: CUMULATIVE DISTRIBUTION OF LOGARITHM OF RATE OF PUBLICATION AT BROOKHAVEN NATIONAL LABORATORY.

Generally, in multiplicative processes, the total probability of the phenomenon is given by the product of each of the probabilities of the subtasks or factors involved in the process, $P = p_1 * p_2 * \ldots * p_n$. Mathematically, this means that small variations in the probabilities of the factors can have an enormous impact in the total probability of the event. Likewise, if the probability of a certain subtask is 0, the multiplicative process will stop altogether. **That is, in a multiplicative process, the loss of a single subtask causes the entire process to fail.** Multiplicative relationship of interdependent events leads to lognormal distributions (West and Deering 1995). This implicit interdependence of the multiplicative processes explains the long tail of their lognormal distributions. The multiplicative nature of the process that shows lognormal distribution is related to what it has been called “the law of proportionate effect” (West and Deering 1995).

5.1.3. The Generative mechanism of Lognormal Distributions in Nascent Entrepreneurship: Theoretical and Practical Implications

Proportionate differentiation has been identified as the generative mechanism of a lognormal distribution (Gibrat, 1931; Mitzenmacher, 2004). Applying this generative mechanism to nascent entrepreneurship, it can take the following formulation. Proportionate differentiation processes have two main components: the initial value and the accumulation rate. Initial value is the amount of a variable that an entrepreneur has at the beginning of the entrepreneurial process, what in the entrepreneurial literature has been described as the initial capital (human capital, financial capital, strong and weak networks, etc.). Accumulation rate here should be considered as the "entrepreneurial performance rate": the rate at which an entrepreneur can increase the amount of an outcome variable in a period of time, such as number of entrepreneurial activities, investment capital, number of clients, number of employees, revenues, business partners, etc. Therefore, according the proportionate differentiation, nascent entrepreneurs differ in outcomes because of both their difference in their inputs at the beginning of the entrepreneurial process (initial entrepreneurial capital) and also their differences in the rate at which they can perform the entrepreneurial activities and aims (entrepreneurial performance rate).

TABLE 4 - PROPORTIONATE DIFFERENTIATION IN NASCENT ENTREPRENEURSHIP

Proportionate differentiation in nascent entrepreneurship	
Initial value	Entrepreneur's inicial capital (human, financial, etc.)
Accumulation rate	Entrepreneurial performance rate (rate at which the entrepreneur can increase an outcome variable (number of clients, investment capital, revenues, employees, etc.)

Given that this is a multiplicative process, the initial value of the input variables (entrepreneurs' initial capital) and their entrepreneurial performance rate interact in a multiplicative way. Future amounts of the outcome (revenues, number of employees, etc.) will depend of **both** the initial entrepreneur's capital **and** the entrepreneurs' performance.

This process can be explained with an example: **Entrepreneur A,** because of his/her family origin, has a remarkable initial entrepreneurial capital: family money, good industry networks, a good team of employees, etc. **Entrepreneur B**, of a modest family origin, has not those high initial resources, but he/she shows a remarkable entrepreneurial performance (ability to search for business opportunities, market insights, motivation, commitment, etc.). Given that the two components are playing a role in the nascent venture, they may later lead to large differences – heavy tailed distribution - in the entrepreneurial outcomes of both entrepreneurial projects. **Entrepreneur A** may fail or may have lesser outcomes because of less entrepreneurial performance rate than **Entrepreneur B**. Or just the

opposite, **Entrepreneur B** may fail or have lesser outcomes than **Entrepreneur A** because of not having enough initial resources. Because entrepreneurs diverge not only on the initial amount of resources (financial, human, etc.) but also on the accumulation rate (here denominated "entrepreneurial performance"), future amounts of outcomes (revenues, number of employees, etc.) would increase greatly for some entrepreneurs, creating a heavy right tail in the distribution. Or, at the contrary, for many other entrepreneurs, future amounts of outcomes would remain at low level, possibly creating a bell-shape head in the distribution (Gabaix, 1999).

This type of interaction between the initial entrepreneurial capital and the entrepreneurial performance is not obvious or trivial. We will see later that there are many organizational processes in which only one of these parameters plays the relevant role in the emergence of the distribution.

Let us consider that, regarding nascent venture outcomes, instead of having the proportionate differentiation as the generative mechanism, we are dealing with pure power law distributions, and, therefore, the generative mechanism may be self-organized criticality (Newman, 2005; Andriani and McKelvey, 2009; Boisot and McKelvey, 2011). In this case, entrepreneurs differ on the value on an outcome because, after some entrepreneurs reach a critical state, some specific events trigger the increase of their entrepreneurial outcomes (revenues, etc.) ranging from small to very large. So, a small event in the entrepreneurial process may produce an "avalanche" that may change completely the ranges of the outcomes (Bak, 1996). In self-organized criticality processes, in order to have large differences in the outcomes, it is required to reach a "critical state". However, this is not necessary in a proportionate differentiation process. Outcome values are a function of the **products of probabilities** between the initial entrepreneurs' capital (human, financial, etc.) and their entrepreneurial performance, as in multiplicative phenomena (see above section "Multiplicative processes"), as it was described above in the

examples of income distribution and research papers production (see previous section). High and extreme values in a pure power law distribution in an entrepreneur's outcomes after reaching a critical state would be unpredictable, or nondeterministic (Bak, 1996; Boisot and McKelvey, 2011; Sornette and Ouillon, 2012), whereas, in a log-normal distribution, these outcomes would be predicted as long as we know the components values, the initial inputs (entrepreneur's initial capital) and the entrepreneur's performance, and their approximate probabilities.

Proportionate differentiation allows the possibility that even if an entrepreneur has high initial resources because of luck or family origins - or any other reason -, the entrepreneurial outcomes may be eventually surpassed by another entrepreneur with a superior entrepreneurial performance. Random differences among entrepreneurs at the beginning of the nascent entrepreneurial process in terms of resources not always lead to eventually long-term differences in nascent venture outcomes (Mankiw, 2013). But proportionate differentiation also allows the opposite: that is, the possibility that an entrepreneur, with superior entrepreneurial performance, may not be never able to catch up the entrepreneur with very large initial resources over time, depending of the probabilities of the different factors that influence the nascent entrepreneurial process.

From a practical point of view, proportionate differentiation also offers us some clues about the way that institutions – governments, business incubators, venture capital, etc. - should allocate the resources in order to promote and foster entrepreneurship and to retain the best entrepreneurial projects. The future of a nascent entrepreneurial project depends on the product between the entrepreneurial performance **and** the initial entrepreneurial resources (Boolean operator "AND"). Therefore, the allocation of resources across different nascent entrepreneurial projects should prioritize those with both higher initial resources and higher entrepreneurial performance. Although this recommendation seems

obvious *prima facie*, the analysis of other generative mechanisms will show below that if the nascent entrepreneurial processes do not follow a lognormal distribution, the allocation of resources should focus only in one of these two elements, initial entrepreneur's resources **or** entrepreneur's performance (Boolean operator "OR").

The next implication for the allocation of resources among nascent entrepreneurial projects - because of the mechanism of proportionate differentiation - is not only to keep large disparity in the allocation of those resources between the best entrepreneurial projects and the ordinary ones, but also among the best performers. The lognormal distribution allows very large entrepreneurial outcomes differences (Joo, Aguinis and Bradkey, 2017).

5.2. EXPONENTIAL TAIL DISTRIBUTIONS

The second plausible distribution in entrepreneurial outcomes worldwide may be the exponential tail distribution. It can take two forms: a pure exponential distribution, or a power law distribution with an exponential cut-off. Both distributions, with positively skewed tails, decay at an exponential rate.

According to our analysis of the international entrepreneurial data sets, these exponential tail distributions may take the form of a power law with an exponential cut-off, consisting in an initial long and heavy head, and then increasingly a falling right tail. It has two parameters, alpha (α) (>1) and lambda (λ) (>0), both rates of decay, that indicate the rate of falling of the right tail. The heaviness of the right tail is determined by the value of the two parameters: α closer to 1, and λ closer to 0 will generate a heavier right tail. Thus, changing the two parameters we can generate a heavy tail such

as one of a lognormal distribution or a very light tail such as an exponential distribution.

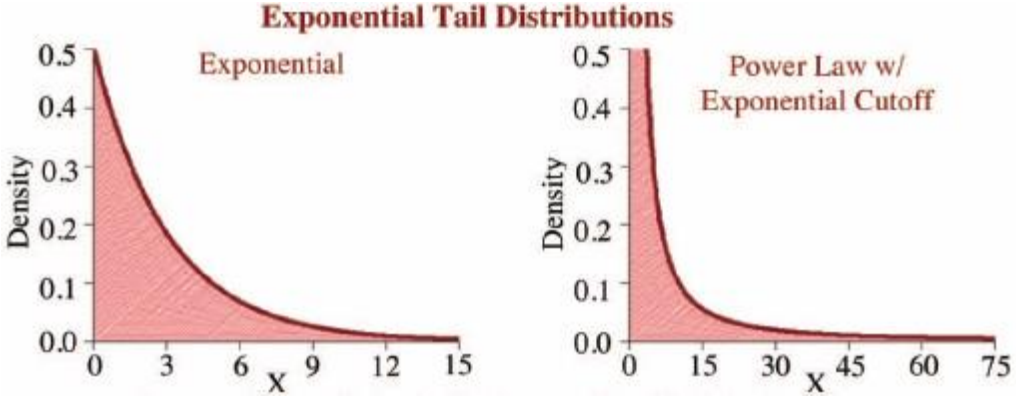


FIGURE 17 - FIGURE FROM JOO, AGUINIS 2017, P. 1024. EXPONENTIAL TAIL DISTRIBUTIONS: EXPONENTIAL (Λ = 0.5), POWER LAW WITH AN EXPONENTIAL CUTOFF (Α = 1.5, Λ = 0.01).

[exponential (λ = 0.5), power law with an exponential cut-off (α = 1.5, λ = 0.01)]

Incremental differentiation is the generative mechanism of exponential tail distributions (Amitrano, 2012; Joo, Aguinis and Bradkey, 2017). According to this generative mechanism, the difference among entrepreneurs in terms of outcomes would be based on their differences with respect to their entrepreneurial performance and **only** on the entrepreneurs' performance ("accumulation rate on the outcome"). This "entrepreneurial performance" refers to the amount of the outcome variable (revenues, number of employees, number of entrepreneurial activities, etc.) that an entrepreneur is able to generate in a time period (revenues per year, number of new clients, venture capital rounds, etc.).

Incremental differentiation is different from proportionate differentiation (the lognormal distribution's generative mechanism). In proportionate differentiation (lognormality) the entrepreneurial outcome variables are explained by **both** entrepreneur's initial resources and entrepreneurial performance. However, in incremental differentiation (exponentiality) the value of the outcome is a function **only** of the

entrepreneur's performance, and only of this, without having to take into account the initial entrepreneur's resources.

The second difference with proportionate differentiation (lognormality) is that incremental differentiation allows the introduction of the "diminishing returns effect" in those entrepreneurs with high entrepreneurial performance through the "exponential cut-off". Entrepreneurs with superior entrepreneurial performance, ultimately, will face steep difficulties as they reach their full capacity. Those entrepreneurs who accrue outcomes more quickly than others would eventually have to face diminishing returns, opening the possibility of generating a distribution of cumulative outcomes that follows a power law with an exponential cut-off (Joo, Aguinis and Bradley, 2017, p. 1032). This process does not require declining entrepreneurial performance over time, but rather the increasing difficulty of getting additional outcomes when the highest levels have been reached ("diminishing returns"). This situation generates a power law with an exponential cut-off (Amaral et al., 2000).

Incremental differentiation as a generative mechanism suggests that entrepreneurs differ in the outcomes because of their entrepreneurial performance, which produces linear increases in outcomes, and tend to have linear effects on their entrepreneurial outcomes rather than multiplicative effects. The differences in the amount of outcome among entrepreneurs exist because some entrepreneurs, compared to others, generate larger increments in outcomes, and, furthermore, entrepreneurs with the highest entrepreneurial performance may have to face diminishing returns.

From a theoretical point of view, if nascent entrepreneurial outcomes are defined by this generative mechanism - incremental differentiation -, we can describe each entrepreneur just in terms of his/her idiosyncratic

entrepreneurial performance. The value of a future outcome value is dependent on the entrepreneurial performance but not on the entrepreneur's initial resources.

From a practical perspective, to generate greater overall entrepreneurial outcomes (increasing revenues, number of employees, etc.) this framework would implement a heavy investment on entrepreneurs with higher entrepreneurial performance than others. Incremental differentiation would recommend allocating resources (venture capital, business incubators, grants, etc.) variably across entrepreneurs - rather than similarly - based on entrepreneurial performance: the nature of past outcomes of an entrepreneur in terms on his/her entrepreneurial performance will determine future outcomes.

5.3. CONCLUSIONS

This research was done in the context of the on-going dialog and debate regarding the search for the generative processes in nascent entrepreneurship, and more broadly, in the discovery of heavy-tail distributions in inputs and outcomes variables across different nascent entrepreneurial panel studies performed in different countries and continents (Andriani & McKelvey, 2009; Reynolds and Curtin, 2011; Crawford and McKelvey, 2012; Crawford et al., 2014; Crawford et al., 2015; Reynolds, 2017,b).

Studying the variables of the American panel (PSED II), Shim (2016) proposed that the lognormal distribution may be one of the best plausible models to describe the long-tail distributions in entrepreneurship and he suggested the multiplicative process as the generative mechanism. Our results deepen and continue his research showing that the lognormal

distribution may be the best fit for American entrepreneurial datasets and probably also for other panels worldwide, and that its generative mechanism, proportionate differentiation, can explain the shape of the distributions.

From a theoretical point of view, our research suggests the confirmation of the presence of a prevalent generative mechanism in the observed outcomes distributions in nascent entrepreneurship: that is, proportionate differentiation. However, our results were not conclusive, and the study of other international panels may be required to confirm our results. With several of the analysed variables, the high *p* value does not allow us to reject the null hypothesis, when comparing lognormal distributions with power law distribution with an exponential cut-off (both type of distributions may be a good fit). Until what extent we need not to consider other generative mechanisms, such as incremental differentiation, will depend on the analysis of other international outcomes datasets. Unfortunately, only four longitudinal panels are now in the public domain (Australia, Sweden, and U.S. PSED I & II), and the harmonized dataset among some of the rest of the projects do not include the outcomes variables studied here (Reynolds et al., 2016; Reynolds, 2017b).

However, from the results, we can infer that, probably, given the outcome differences among nascent ventures, the generative mechanisms based on homogenization (Normal, Poisson, Weibull) are not adequate. On the other hand, datasets do not show an increase at an explosive (nonlinear) rate, and, therefore, we can also discard generative mechanisms such as self-organized criticality – pure power law distributions - (Andriani and McKelvey, 2009). The disregard of mechanisms such as incremental differentiation – power law with an exponential cut-off - will require further empirical research on other countries datasets. The results of this research also suggest that the emergence of a nascent venture is not merely

> *"like entering into a lottery (…) with high death rates, skewed returns with most players losing out, random growth, little or no entrepreneurial learning (…), no influence of education on performance, little control over outcomes" (Nightingale and Coad, 2014, p. 130).*

Indeed, the different probabilities of the factors that make successful a nascent venture play a major role, but it works following a determined generative mechanism, not simply by rolling a dice (Nightingale and Coad, 2014, p. 130).

Although previous research pointed out to power law distributions as the prevalent in nascent entrepreneurship, at that time, there were less sophisticated statistical packages to implement accurate distribution pitting. This would have led to certain uncertainty about the better fit among heavy-tailed non-normal distribution (Crawford et al., 2015; 2016). At this point, our results (and of Shim's, 2017) suggest that the pure power law distribution and its generative mechanism, self-organized criticality, might not be completely suitable for explaining nascent entrepreneurial outcomes distributions - or at least, major sections of the complete empirical datasets - although it may be needed to explain other aspects of the entrepreneurial process (Andriani and McKelvey, 2009; Crawford et al., 2015; Joo, Aguinis & Bradley, 2017).

As mentioned above, the empirical datasets on nascent entrepreneurship shows distributions in which it is difficult to conclude if they should considered as lognormal or power laws. Also, there are certain parts of the distributions that are better fitted by lognormal, and another parts of the distributions that are better fitted by a power law distribution.

Initially, Crawford and McKelvey (2012) identified the outcomes of the longitudinal panels as power laws; subsequently, using more powerful and precise software, Shim (2016) was able to identify the lognormal

distribution as a better fit for the entrepreneurial data. However, this should not come out as a surprise because there is a robust relationship between these two distributions, between lognormality and inverse power law distributions (West and Deering 1995, p. 156). As a system, functioning in lognormal mode, become more and more complex, its distributions become broader, increasing the value of the variance, and it starts to take characteristics related to a system that show power law distribution patterns, such as scale-invariance or fractality (West and Deering 1995, p. 156).

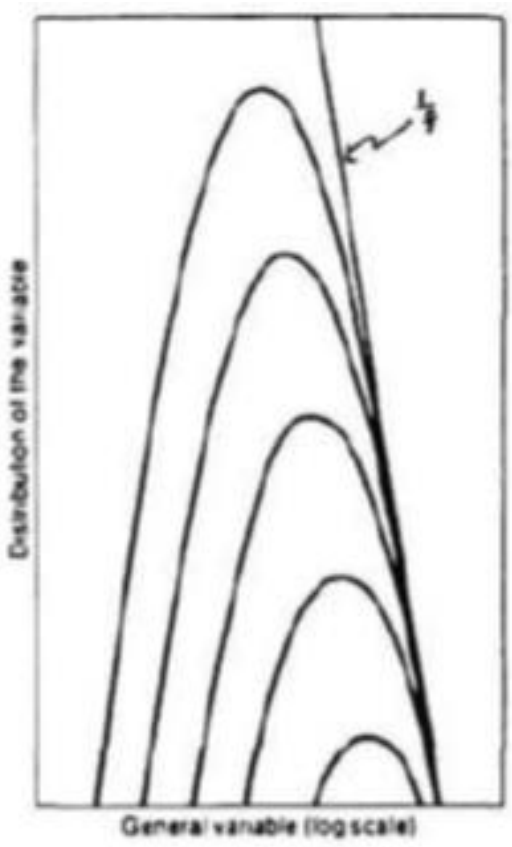


FIGURE 18 - FIGURE TAKEN FROM WEST AND DEERING (1995, P. 157) BASED ON KOLMOGOROV CLASSICAL ARTICLE (1941). AS A LOGNORMAL DISTRIBUTION BECOME BROADER, WITH A HIGHER VARIANCE, CORRESPONDING TO AN INCREASE IN THE COMPLEXITY OF THE SYSTEM, THE DISTRIBUTION RESEMBLANCES AN INVERSE POWER LAW (THE STRAIGHT LINE IN THE FIGURE). A VERY COMPLEX LOGNORMAL PROCESS TAKES ON MORE THE CHARACTERISTICS OF AN INVERSE POWER LAW DISTRIBUTION.

The more subtasks have to be realized to implement the process, the greater is the range of values of the variable, and the lognormal and power law distributions become indistinguishable. That is, if the lognormal

distribution has a large variance, which corresponds to a process with a large number of subtasks, and we take a sample in the region of the higher values of the variable, it will be very difficult to discriminate between a lognormal distribution and a power law distribution (West and Deering 1995, p. 160; Fig. 17). If the number of required subtasks to complete the process increases, the distribution changes from a lognormal to a power law distribution.

Similarly, as the complexity of the process increases along with large number of subtasks, the slope of the distribution decreases: smaller slopes point out to more complex processes. As the slope of the distributions increases, the complexity decreases - smaller number of subtasks - (West and Deering, 1995, p. 179 and ff.). The increment in the slope of the distribution points out to a reduction in the complexity of the process. Therefore, applying this to nascent entrepreneurship, the outcome variables distributions of a country that requires many subtasks to start-up a new venture – such as bureaucracy, venture capital sector poorly developed, etc. - should show a distribution resembling a power law and a smaller slope (higher complexity) rather than a log normal approximation (West and Deering, 1995).

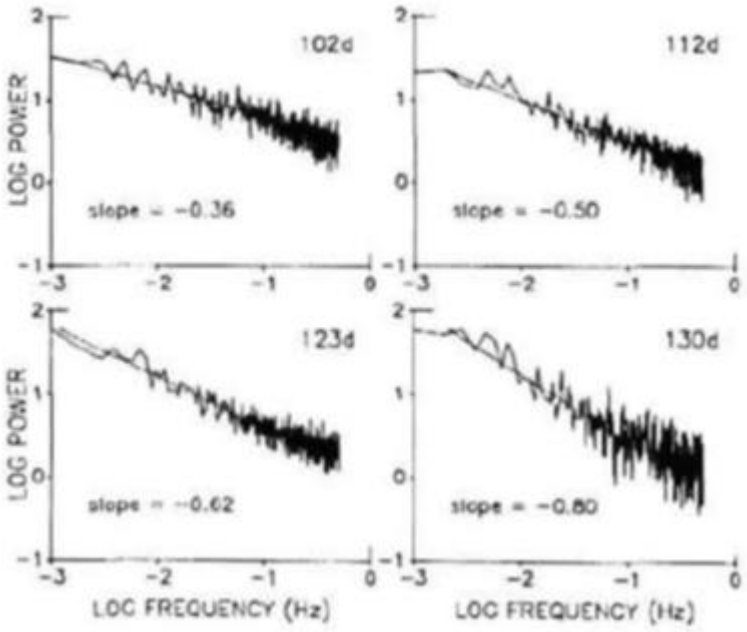


FIGURE 19 - FIGURE 3.21 TAKEN FROM WEST AND DEERING (1995, P. 160): EXAMPLE (BREATHING RATE IN FETAL MATURATION) OF A PROCESS IN WHICH THE DECREASE IN COMPLEXITY GOES TOGETHER WITH AN INCREMENT IN THE SLOPE OF THE SPECTRUM (THE SLOPE CHANGES FROM -0.36 TO -0.80).

If the lognormal distribution and its generative mechanism - proportionate differentiation - is indeed relevant for explaining nascent entrepreneurial outcomes distributions, then, we can conclude the presence of **positive feedbacks** between past and future entrepreneurial initial resources and outcomes. Proportionate differentiation implies that both initial resources and the entrepreneurial performance will determine the future outcomes of the nascent venture, and that both aspects will affect them. Proportionate differentiation may allow “outcome loops”: some entrepreneurs can receive larger outcomes increases because positive feedback between past and future outcomes. For example, an initial large number of clients would attract venture capital to initiate a new and/or bigger financial round. This may lead to some explosive, non-incremental increases in the new venture outcomes. By “amplification events”, the multiplicative nature of lognormal processes can go through a transitional period into a power law distribution (West and Deering, 1995). As

mentioned above, inverse power-law distributions are also observed as the result of multiplicative processes that show multiple amplification events.

A classic example of this transition from lognormality to power law distribution through positive loops can be observed in the Pareto's distributions of income. It is relevant to notice that the lognormal and the inverse power law distribution coincide for much of the dataset, and therefore, both distributions would be a valid fit for most of the data range. The lognormal distribution is the better fit for most of the dataset, except in the last percentiles, corresponding to those in the society with highest income, in which the distribution changes into an inverse power law. West and Deering (1995) suggested that this transition may help to understand the nature and the mechanisms that explain the distribution of population income. For the majority of the population, the generative mechanism of proportionate differentiation applies and the lognormal distribution holds. However, for the richest, the income, somehow, is amplified: the process of earning money shows a peculiar mode that allows the accumulation of extra wealth by some kind of amplification processes that lead to the emergence of a power law distribution, "the rich get richer" (Mathew effect) (West and Deering, 1995 p. 172). The creation of a new venture may be one of the modes of amplification, because the initial forms of capital can be amplified through the efforts of others – the employees, clients, suppliers, etc .-, value creation, financial leverages, investment returns, etc. (West and Deering, 1995). Similar amplifications with their positive loops effects may also occur among different nascent entrepreneurial projects. For most of them, the lognormal distribution may hold, but for the super-entrepreneurs, at the extreme of the heavy tail of the distribution, explosive growth in outcomes may also happen, becoming power law distributions.

Concluding: the nascent entrepreneurial process is therefore different than many of the individual output samples analysed by Joo, Aguinis and Bradley (2017), which shows incremental differentiation as

generative mechanism. In those cases, future individual outputs are determined only by the accumulation rate (the “performance”), and not by initial outputs. In nascent entrepreneurship, however, the initial entrepreneur’s resources will also affect profoundly future outcomes of the nascent venture.

On the other hand, if our conclusions are correct, high variability among nascent ventures outcomes would produce even higher variability among them in those outcomes in the future. In a process ruled by homogenization generative mechanisms (that is, showing symmetric distributions such as the normal) higher variability in the past will be followed by lower variability among the outcomes in the future. In a process ruled by incremental differentiation high variability across individuals will produce higher variability in individual outputs in the future but only in terms of output accumulation rate differences among individuals. Whereas, if the process is proportionate differentiation, such as nascent entrepreneurship, the future variability in new ventures outcomes will also increase, although depending not only on entrepreneurial performance but also on entrepreneur’s initial resources.

6. AGENT-BASED MODELLING AND SIMULATION (ABMS) IN ENTREPRENEURSHIP

6.1. ABMS SOFTWARE AND TOOLKITS

Repast Symphony (Repast, 2017), NetLogo (Wilensky, 1999), or MASON (2016) are examples of special-purpose agent tools, “Dedicated Agent-based Prototyping Environments” that provide the user with special features focused on ABMS. This research uses NetLogo, which is a free ABMS environment developed at Northwestern University’s Center for Connected Learning and Computer-Based Modelling (Wilensky, 1999) [http://ccl.northwestern.edu/netlogo/resources.shtml].

NetLogo uses a modified version of the Logo programming language and it consists in a graphical environment with mobile agents - called “turtles” - that reside in a world of “patches” - as square grid cells -. The environment and agents are observed and monitored by an “observer.” “Primitives” are the Netlogo programming language’s built-in commands. NetLogo also includes a participatory “HubNet”, in which users can upload share and discuss the models.

General-purpose desktop computational mathematics system such as MATLAB or *Mathematica* can also be used to develop agent-based models (North and Macal, 2007). In this research, the statistical software R (R Core Team, 2018) has been used, in conjunction with NetLogo, to investigate the heavy-tailed distributions in the datasets produced by our model by NetLogo simulations. Stochastic agent-based simulations can generate such amount of data that large-scale softwares for data analysis

are required to process that information (Thiele et al, 2014; ten Broeke et al, 2016).

Our “Nascent Entrepreneurial Agent-based model” is coded in Netlogo, one the most common software platform for ABMS. Two of the most relevant and standard textbook on ABMS (Railsback and Grimm, 2012, new ed. 2019; Wilensky and Rand, 2015) use Netlogo as toolkit. Netlogo has a professional design, comprehensive documentation, high-level programming language, with many built-in commands (“primitives”), integrated graphical user interface, integrated tool for performing simulation experiments (“BehaviourSpace”), and a very active user community. In The CoMSES Net Computational Model Library the main global repository of ABMS, the majority of the models are implemented in Netlogo (Railsback et al., 2017) (http://www.openabm.org/models).

For the development of the Agent-based model, Railsback and Grimm (2012) propose an iterating “modelling cycle”, where iteration is established as a method of continuous improvement of the model. Railsback and Grimm’s modelling cycle has the following steps:

1. Formulate the research question.
2. Assemble hypotheses for essential processes and structures, starting just with the minimum number of the model factors.
3. Choose scales, entities, state variables, processes, and parameters.
4. Implement the model.
5. Analyse, test, and revise the model.

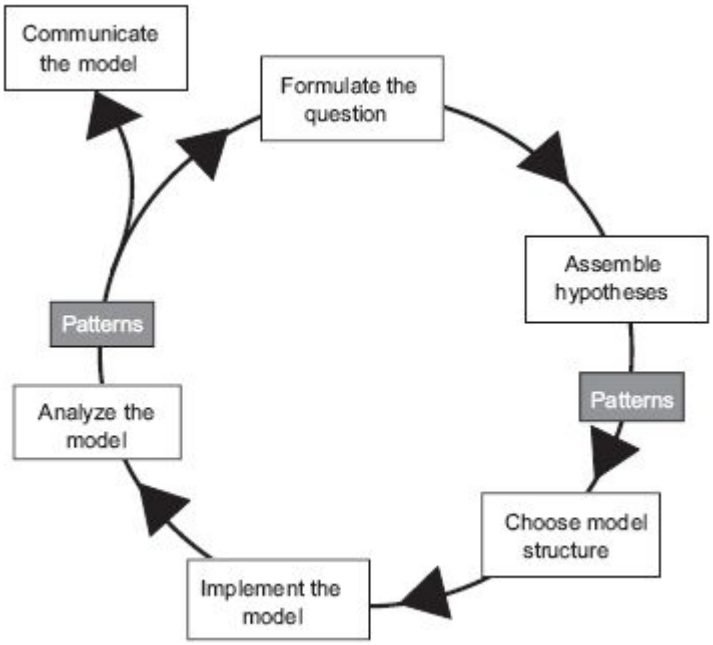


FIGURE 20 - RAILSBACK AND GRIMM'S MODELLING CYCLE (*FROM RAILSBACK AND GRIMM, 2012, P. 7-9.)*

One of the most important tasks in developing an agent-based model is the identification of the agent types and the definition of their attributes. Secondly, the agent behaviour has to be specified using a theory of agent behaviour or a behavioural heuristic. And third, the ways of interaction among the agents is added – which, when and how they interact - (Macal and North, 2009). Marcal and North have already explored some of the special characteristics of these agents, their behaviour and interactions in the business and management domain (North and Macal, 2007).

A key figure of agent-based modelling is how to model agent relationships and the dynamics that governs those interactions. Marcal and North (2009) have described some of the most common topologies used in ABMS for representing social interactions (see figure 21 below):

-a. The aspatial model, in which agents do not have location and the model is not associated with a particular space representation or area.

-b. Cellular automata, in which agent interactions are patterned based on a grid or lattice.

-c. The Euclidean space models, in which agents wander in two or more dimensional spaces.

-d. Geographic Information System (GIS) topology, in which the agents move on a realistic geo-spatial landscape.

-e. Network topology, a collection of nodes connected by links, that can be static or dynamic.

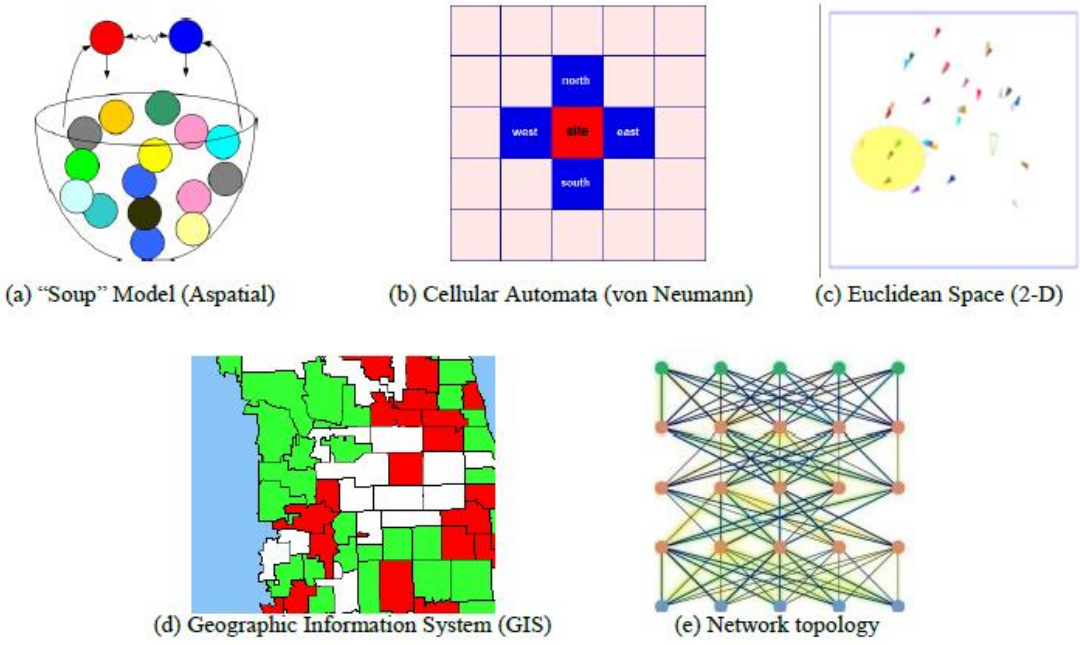


FIGURE 20 - ***TOPOLOGIES FOR AGENT RELATIONSHIPS AND SOCIAL INTERACTION, FROM MACAL AND NORTH (2009, P.94).***

In any of these topologies, there are local interactions and local information transfers between agents, with a limited connectivity and in which information is confined to local exchanges. Here "local" does not have to be spatial proximity: for example, network topology allows that the

agents to be linked by relationships rather than by physical proximity. On the other hand, there is neither a global controller, nor access to global information by the agents (Macal and North, 2009).

6.2. DESCRIPTION OF THE MODEL

6.2.1. THE ODD (OVERVIEW, DESIGN CONCEPTS, DETAILS) PROTOCOL

In order to describe the model, this project will follow the protocol ODD developed by Grimm et al. (2006, 2010) (The protocol ODD is part of the document "TRACE", see section below). This protocol creates a standard structure and a generic format to document ABMS, facilitating the completeness of the model description and an easier way to replicate them (Railsback and Grimm, 2012). It also reassures that the theoretical background and assumptions of the model is clearly stated (Grimm, 2010). Grimm's protocol (2010) is defined by seven elements, with the following sequence (Grimm et al., 2010, Table 1, p. 2763):

1. Purpose.
2. Entities, state variables, and scales.
3. Process overview and scheduling.
4. Design concepts:
    a. Basic principles.
    b. Emergence.
    c. Adaptation.
    d. Objectives.
    e. Learning.
    f. Prediction.
    g. Sensing.

    h. Interaction.
    i. Stochasticity.
    j. Collectives.
    k. Observation.
5. Initialization.
6. Input data.
7. Submodels.

A detailed description of the elements of protocol can be downloaded from Railsback and Grimm's (2012) web companion here:

http://www.railsback-grimm-abm-book.com/Chapter03/GrimmEtAl2010_ODD-Update-1.pdf

The protocol template, here:

GrimmEtAl2010_App2_ODD-template.docx

6.2.2. Concepts For The Evaluation of Agent-Based Models

A computer model, especially an agent-based model, should be able to answer questions related to the system that we are studying: it must provide outputs relevant to the model user (Wilensky and Rand, 2015). The accuracy of a model is traditionally evaluated through three processes: **validation**, **verification**, and **replication**.

Wilensky and Rand (2015) define these processes in the following way:

- "**Model validation** is the process of determining whether the implemented model corresponds to, and explains, some phenomenon in the real world."
- "**Model verification** is the process of determining whether an implemented model corresponds to the target conceptual model".
- "**Model replication** is the implementation by one researcher or group of researchers of a conceptual model previously implemented by someone else". (Wilensky and Rand, 2015, p. 311-2).

However, many agent-based models have a stochastic nature, such as the one we are introducing here, and, therefore, the methodologies of verification, validation, and replication have to rely in statistical methods, through multiples runs, in order to confirm the accuracy of a model. On the other hand, across disciplines, terminology is ambiguous and it is difficult to elucidate the processes of evaluation of a model and the clear definition of these accuracy terms (Augusiak et al., 2014). That is why many scholar modellers have made a call for reviewing the terminology and to unify the criteria for the evaluation of agent-based models (Grimm et al., 2014; Augusiak et al., 2014). This research will follow the latest trend in model evaluation, called the "TRACE" documentation, but, first, we will define some of the major concepts for the evaluation of an agent-based model.

VERIFICATION

An agent-based model is incremental in nature. Progressively we add parts (and/or remove others), and, as the models grows in complexity,

it become more difficult just looking at its code to figure out if it is performing its function and to understand the conceptual model behind the code. It is a process in which, incrementally, we verify the alignment between the conceptual model and the code.

One way to describe the conceptual model underlying our model is to use flowcharts. Our nascent entrepreneurial agent-based model flowchart is depicted below in the “description of the model” section, and it describes the flow of decisions happening during the operation of the software code. The flowchart diagram is also rewritten in pseudo-code. “Pseudo-code” is as a “midway” point between natural language and formal programming language that can be read for anyone, regardless the reader’s knowledge of Netlogo programming language, facilitating the verification process (Wilensky and Rand, 2015). In this document, pseudo-code has also been used in some descriptions of the ODD (Overview, Design concepts, Details) protocol (see below).

The process of verification also focuses in the elimination of “bugs” from the code in order to guarantee that it follows the conceptual model. Sometimes, what it seems a “bug” is not, but rather an oblivious characteristic of the system. For example, in our tests at the limit, we analysed the model having the global variable “Social dynamism” at 0. With parameter social dynamism = 0, some “entrepreneur-opportunities” entities appeared in the world in some runs, when, theoretically, they should not. However, in this case it was not a *bug*. Given the geographical conceptual framework of the model and the random location set-up, some of the opportunities may coincide in the same spatial patch with an entrepreneur at the very beginning of the run, and, therefore, if their state variables match, they become an entrepreneur-opportunities entity.

In complex agent-based model, such as this one, verification can be very difficult to achieve. A strange result may be the product of a bug in the code, an error in the translation of the conceptual model into programming code, or an unexpected outcome that emerge from the nature of the agents' interactions. Therefore, a model is not either verified or unverified: it exists along a continuum of verification. It is always possible to check more in detail the parameters space, or to write more component tests or to make more sensitivity analyses (Wilensky and Rand, 2015).

Sensitivity, Uncertainty and Robustness Analysis

In agent-based modelling, sensitivity analysis explorers how sensitive the model is to the set of initial conditions, how sensitive the outputs of the model are to changes in parameters values (Thiele, Kurth and Grimm, 2014). This procedure implies to vary the group of parameters of the model, or to add new parameters into the model and to study the variations in the results. **Parameters** are the constants in the Netlogo's primitives, equations and algorithms that are used to represent the processes in an agent-based model. ***Parameterization*** is the task of selecting values for the parameters of a model to relate it to real system as much as possible (Railsback and Grimm, 2012). "Direct parameterization" is when parameter values are obtained directly from the literature or experts, and "inverse parameterization", when we define parameter values inversely by calibrating the model to real data (Grimm et al., 2014).

> *"Sensitive analysis is an examination of the impact of varying model parameters on model results" (Wilensky & Rand, 2015, p. 23).*

Thus, we analyse the effect that initial conditions and agent mechanisms have on model results. We can distinguish between "local sensitivity analysis", which is performed one parameter at a time, sweeping

parameters and collecting multiple runs, and “global sensitivity analysis” – much more complicated computationally -, in which several or all parameters are varied over their whole ranges (Grimm et al., 2014). The amount of data generated by this parameter sweeping can be so large that requires specific big data tools to study the results. In our case, we have used R (R project, 2018), a common current software in statistics and data sciences (Thiele et al, 2014; ten Broeke et al., 2016).

However, sensitivity in the quantitative results does not necessary mean there will be sensitivity in the qualitative results. As we will see below, this is the case of our nascent entrepreneurial model, in which, although the distributions at the end of the run may be different if we change the range of parameters, they are mostly heavy-tailed distributions (power laws, log-normal, Weibull), and scarcely Gaussian (normal ones). This behaviour is probably due to the multiplicative processes that undergone the agents throughout the run. It may also be relevant to study the environmental parameters in which the model operates. Our model uses a two-dimensional torus grid of a certain dimension. The area of this grid and the global environmental variables (“social-dynamism”) affect greatly the wandering of the agents in this world and the model results.

Uncertainty analysis explores how uncertainty in parameter values affects the reliability of the results of the model (Railsback and Grimm, 2012). Although it uses similar techniques of those of sensitivity analysis, the objective is different. It aims to understand how the uncertainty in parameter values and the model’s sensitivity to parameters interact to cause uncertainty in the results of the model (Railsback and Grimm, 2012). In many occasions, the value of several parameters in the agent-based model are uncertain for different reasons, for example, because it is a simplification of a process that is not so simple or constant, or its value has not been measured precisely, or – simply - we do not have all the parameters of a real system. However, although parameter uncertainty may

cause high uncertainty in absolute terms, other important results – such as relevant patterns - can be much less affected by parameter uncertainty. Thus, although a simulation model may have uncertain parameter values, it still can be very useful when we use them for *relative* predictions.

Robustness analysis explores the robustness of results and conclusions of a model to changes in its structure (Railsback and Grimm, 2012). Sensitivity analysis often focuses on the response of the model to *small* changes in the values of the parameters. In contrast, robustness analysis focuses more on the response of the model to *drastic, radical* changes in the structure of the model. Robustness testing explores the limits of the model such as setting the parameters to the minimum – or maximum - (extreme values), and forcing the model ("stress tests", "limit tests"). Underlying this technique is the idea that if the ability of a model to reproduce a pattern of the real system is very sensitive to its details, it means that it is not robust and that probably is not able to capture the real mechanisms driving the real system (Railsback and Grimm, 2012).

Robustness analysis is a systematic deconstruction of a model by forcefully changing the model parameters, structure of submodels (simple vs. complex, on/off), and representation of processes (Grimm and Berger, 2016).

VALIDATION

Validation is the process of guaranteeing that there is correspondence between the agent-based model and the real world (Wilensky & Rand, 2015). However, by definition, a model is a simplification of reality. It cannot reflect all of the same features and patterns that exist in the real world. The objective of implementing a model is to incorporate those aspects of the real world that are relevant to our questions. According

to Wilensky and Rand (2015, p. 326), validation has to be considered in two dimensions:

Level of the validation process:

- *Micro-validation*: the behaviour and mechanism encoded into the agents match up with the real world. Micro-validation informs if the model has captured the important parts of the agent's individual behaviour.
- *Macro-validation*: the aggregate, emergent properties of the model correspond to aggregate properties in reality. Macro-validation informs if the model has captured the important parts of the system as a whole.

In other forms of modelling, such as equation-based modelling, only macro-validation is performed. The aggregate results of the equation-based model are compared to the aggregate results of the real system under study. However, agent-based modelling produces results at all level of aggregation.

Level of detail of the validation process:

- *Face validation*: the mechanism and properties of the model look like mechanisms and properties of reality. *Prima facie* (without detailed analysis) the model can convince that it contains elements and components that correspond to agents and mechanisms of the real world. Face validity can exist at both the micro-levels and at the macro-levels of the model.
- *Empirical validation*: the model generates data that correspond to similar patterns of data in the real world. Data produced by the

model must correspond to empirical data of the studied system. Empirical validation, therefore, often implies statistical tests and comparison between data sets. One of the problematic aspects of this type of validation is that real data is frequently with “noise”, difficult to obtain, and partial. On the other hand, reality is not a computational machine with precise and well-defined results, but rather it yields messy results and it is very challenging to isolate and measure the parameters of the real world. Again, empirical validation can be performed at both the micro and macro-levels. In this context, **calibration** is the process of finding the parameters and initial conditions that makes the model to match up as close as possible to the real, empirical datasets.

Calibration is, thus, a special kind of parameterization in which we try to find the best parameters to reproduce patterns observed in the real system (Railsback and Grimm, 2012). Calibration has three purposes:

- To force the model to match empirical results as well as possible.

- To estimate the value of parameters that we cannot evaluate directly. In many real complex systems, there are variables’ values that we cannot know. When we do not know those values, we estimate them “inversely” by adjusting them until the model best matches some observations. This type of calibration is called “inverse modelling” or “inverse calibration”. For several parameters of the real process that we do not have access in our “Nascent entrepreneurial agent-based model” we have made use of this method of “inverse modelling” to define the value - or range of values-.

- To test a model's structural realism. Is it possible to match the empirical results within a reasonable range?

Validation refers to the certainty that the implemented model and agents are similar to reality in those aspects that are relevant for the research questions. Often, many equally good models can be implemented. The key factor is that there is a defensible, reasonable connection between the model and the real world. Railsback and Grimm (2012) have proposed the "pattern-oriented modelling" as another form of empirical validation. A more valid model is achieved when the model is able to match pattern of empirical data at multiple levels. Breig, Coblenz and Pelz (2018) have recently proposed a method to compare simulation outputs of entrepreneurial simulations with the empirical data – "*possible simulation parameter range"* (PSPR) - in order to improve the validation process.

Thus, these five types of validation (micro-face, macro-face, micro-empirical, macro-empirical and pattern-oriented) define the majority of the validation processes. However, an agent-based model will be never a perfect correspondence to reality. The objective of model building is to answer a research question and to explain some results, not to simulate all the aspect of a system (Wilensky and Rand, 2015). Similar to verification, a model is not either valid or invalid. A model is said to be more valid based on how close it is in comparison to the real system. The validation process has also challenging epistemological issues: it assumes that some features in the model correspond to some features in reality. However, are we sure that these features belong to reality? As we established in the first part of this thesis, nascent entrepreneurial empirical data show heavy-tailed distributions in their output. But with current statistical techniques and computational capabilities, we can only determine until certain degree of accuracy which type of distribution they can be. Statistically, one or more distributions can be good fit for the real dataset. Thus, the model has to

mimic – simulate - a blurry, messy and noisy real dataset, which seems, somehow, a puzzling mission.

Agent-based model are frequently stochastic, therefore, they do not produce the same results even given the same initial parameters. This stochastic nature makes more challenging the process of validation and it makes compulsory statistical tools and tests to determine if the model is producing a distribution consistent with that produced in the real system. In our case, for the “nascent entrepreneurial agent-based model”, we will use the R distribution pitting package “**Dpit**” (‘Dpit’ version 1.0: Joo, Aguinis and Bradley, 2017, based in the Kolmogorov-Smirnov test) and the distribution testing R package “**goft**” (package ‘goft’ version 1.3.4: Gonzalez-Estrada and Villasenor-Alva, 2017).

The outputs of the different run can be classified into two types: invariant results and variant results (Brown et al., 2005). Invariant results occur no matter how many times we run the model. Variant results change depending on how the model evolves. In our “nascent entrepreneurial model”, the invariant feature is the statistical persistence of the heavy-tailed distribution at the end of the runs. However, the parameters of these heavy-tailed distributions are different in every run. When the variant results are quite prominent, it may be caused by a path dependent process in the model. A path dependent process is one where the history of the process greatly affects its final state (Wilensky and Rand, 2015). Our model also shows path dependence: some runs, even with the same parameters, may produces very different results.

Replication

As part of any scientific process, replication in computational models has the same relevance than in the subject of physical experimentation. It is defined as:

> *"the implementation by one scientist or group of scientists of a conceptual model (replicated model) described and already implemented (original model) by a scientist or group of scientists at a previous time" (Wilensky & Rand, 2015, p. 337).*

Replication helps to prove that the results are not due to mistakes or omissions, and it increases the model verification since a new implementation of the conceptual model yields the same results than the original. The original model and an associated replicated model may differ across these dimensions such as hardware platform, computer language (Java, Fortran, etc.), toolkits for building the agent-base model (Repast, Ascape, MASON, Netlogo), or algorithms. In any case, besides those differences, a successful replication has to be able to produce outputs sufficiently similar to those of the replicated original. Axtell et al. (1996) have explored the different criterion that should be considered a standard able to judge the level of success of a replication: "numerical identity", "distributional equivalence" and "relational alignment".

6.2.3. The TRACE Documentation ("TRansparent And Comprehensive model Evaludation")

The development of an agent-based model is an iterative process that requires multiples rounds of testing, analysis, and application (Grimm and Railsback, 2005; Schmolke et al., 2010; Grimm et al., 2014; Augusiak

et al., 2014). During this iterative process, several alternative submodes or designs are tested, improved or discarded, as they were introduced in the model at different stages of the model development (Grimm et al., 2014).

The TRACE documentation provides a standard framework for the transparent and comprehensive documentation of models and the underlying modelling process, and it is increasingly adopted in biological and ecological research, for example, in chemical risk assessments of ecosystems (the EU's founded "CREAM project" - http://cream-itn.eu/trace) (Grimm et al., 2009).

This standard protocol ("TRACE") allows:

- The gathering of the whole modelling process: model development, testing and analysis, and application.
- The template for day-to-day documentation of the iterative process and changes and variation of the model/s.
- The facilitation of the organization of the modelling process by modellers.
- The facilitation of the assessment of model quality and suitability by other scholars or decision makers.

The TRACE protocol includes the full model description in the standard format ODD (Grimm et al., 2006, 2010) mentioned in the previous section. It also incorporate the new term 'evaludation' referred to this type of comprehensive quality assessment performed during the TRACE document development (Augusiak et al., 2014). The "evaludation" concept somehow encompass previous terms such as 'validation', 'verification', and 'evaluation', that try to assess if the model is good enough for its intended purpose (Grimm et al., 2014). However, these terms - 'validation', 'verification', 'evaluation', "testing" - can be used in different contexts, they have been interpreted in very different ways through model quality

assessment literature and they do not always capture the iterative nature of agent-based modelling development (Augusiak et al., 2014).

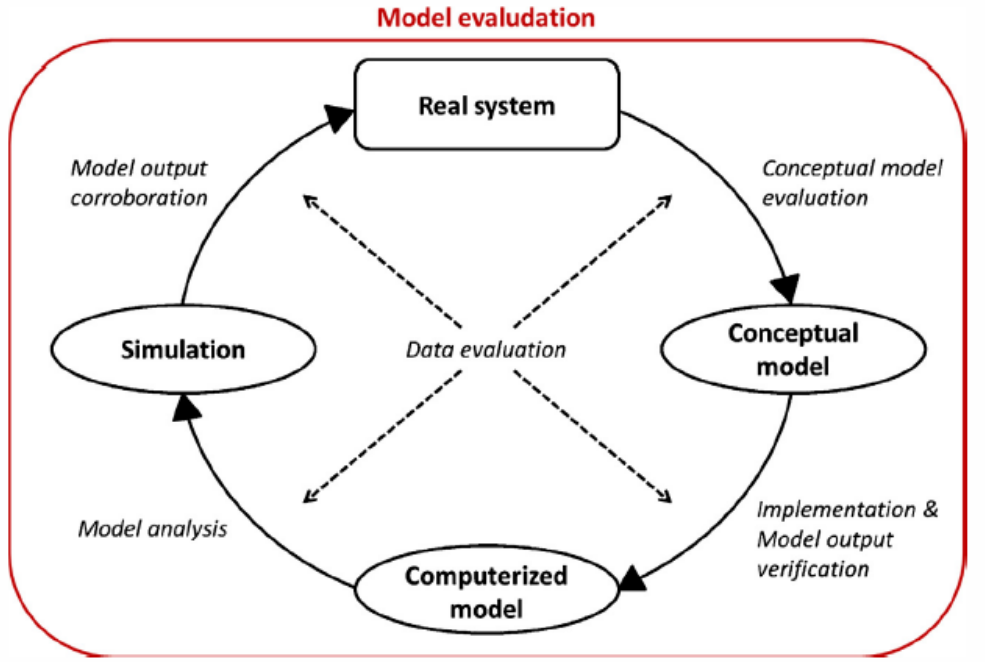


FIGURE 21 - AUGUSIAK'S ET (AL., 2014, P. 5) REPRESENTATION OF THE MODELLING CYCLE. IT HAS FOUR STEPS OF MODEL DEVELOPMENT AND THEIR CORRESPONDING ELEMENTS OF "EVALUDATION".

"Evaludation" -as a methodology- has six elements:

1)) '*data evaluation'*, that assess the quality of numerical and qualitative data used for model development and testing.
2) '*conceptual model evaluation'*, that analyses the assumptions underlying the design of the model.
3) '*implementation verification'*, that checks the implementation of the model (equations and software).
4) '*model output verification'*, that compares the output of the model to the empirical data and patterns that led the design of the model and its calibration.

5) ‘*model analysis’*, that examines the sensitivity of the model to changes in parameters and formulation, in order to understand the key behaviours of the model and the description and justification of the simulation experiments.
6) ‘*model output corroboration’*, that compares the output of the model to data and patterns that were not used for the development and parameterization of the model (Grimm et al., 2014).

The following table describes the structure of the proposed standard for agent-based model descriptions:

TRACE element	This TRACE element provides supporting information on:
1. Problem formulation	The decision-making context in which the model will be used; the types of model clients or stakeholders addressed; a precise specification of the question(s) that should be answered with the model, including a specification of necessary model outputs; and a statement of the domain of applicability of the model, including the extent of acceptable extrapolations.
2. Model description	The model, i.e. a detailed written model description. For individual/agent-based and other simulation models, the ODD protocol is recommended as standard format. For complex submodels, include concise explanations of the underlying rationale. Model users should learn what the model is, how it works, and what guided its design.
3. Data evaluation	The quality and sources of numerical and qualitative data used to parameterize the model, both directly and inversely via calibration, and of the observed patterns that were used to design the overall model structure. This critical evaluation will allow model users to assess the scope and the uncertainty of the data and knowledge on which the model is based.
4. Conceptual model evaluation	The simplifying assumptions underlying a model's design, both with regard to empirical knowledge and general, basic principles. This critical evaluation allows model users to understand that model design was not ad hoc but based on carefully scrutinized considerations.
5. Implementation verification	(1) Whether the computer code implementing the model has been thoroughly tested for programming errors, (2) whether the implemented model performs as indicated by the model description, and (3) how the software has been designed and documented to provide necessary usability tools (interfaces, automation of experiments, etc.) and to facilitate future installation, modification, and maintenance.
6. Model output verification	(1) How well model output matches observations and (2) how much calibration and effects of environmental drivers were involved in obtaining good fits of model output and data.
7. Model analysis	(1) How sensitive model output is to changes in model parameters (sensitivity analysis), and (2) how well the emergence of model output has been understood.
8. Model output corroboration	How model predictions compare to independent data and patterns that were not used, and preferably not even known, while the model was developed, parameterized, and verified. By documenting model output corroboration, model users learn about evidence which, in addition to model output verification, indicates that the model is structurally realistic so that its predictions can be trusted to some degree.

FIGURE 22 - STRUCTURE, TERMINOLOGY, AND CONTENTS OF TRACE DOCUMENTS BASED IN GRIMM ET AL. (2014).

7. THE TRACE DOCUMENT OF "A NASCENT ENTREPRENEURIAL AGENT-BASED MODEL"

7.1. BASIC INITIAL INFORMATION ON THE MODEL

LOCATION

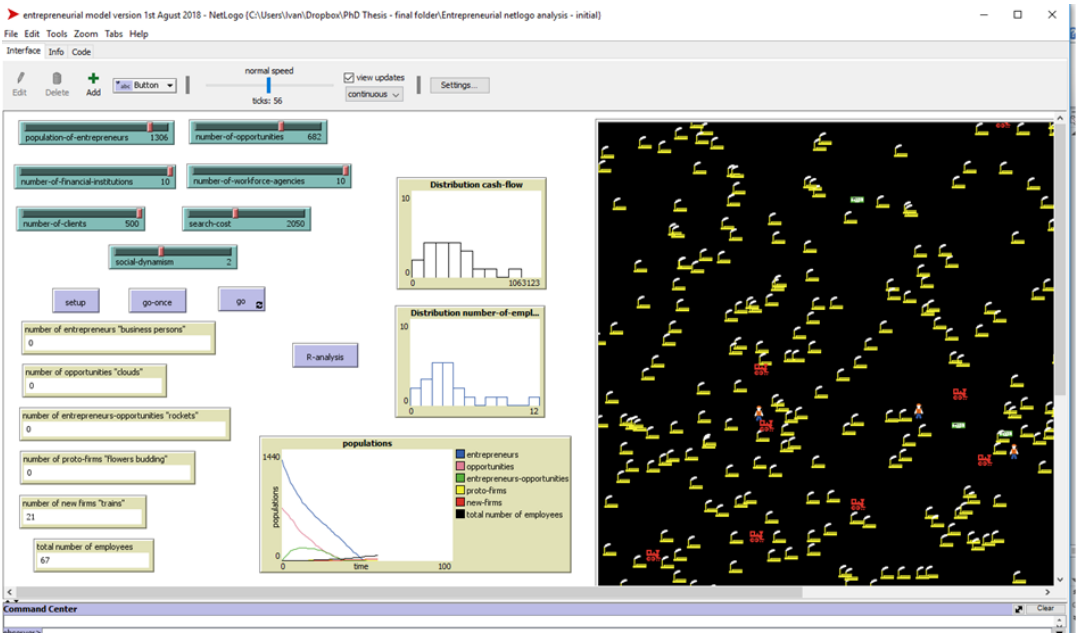
The model will be located at:

- CoMSES Computational Model Library maintained by the OpenABM consortium: (http://www.openabm.org/models)
- Modeling Commons: (http://modelingcommons.org/)

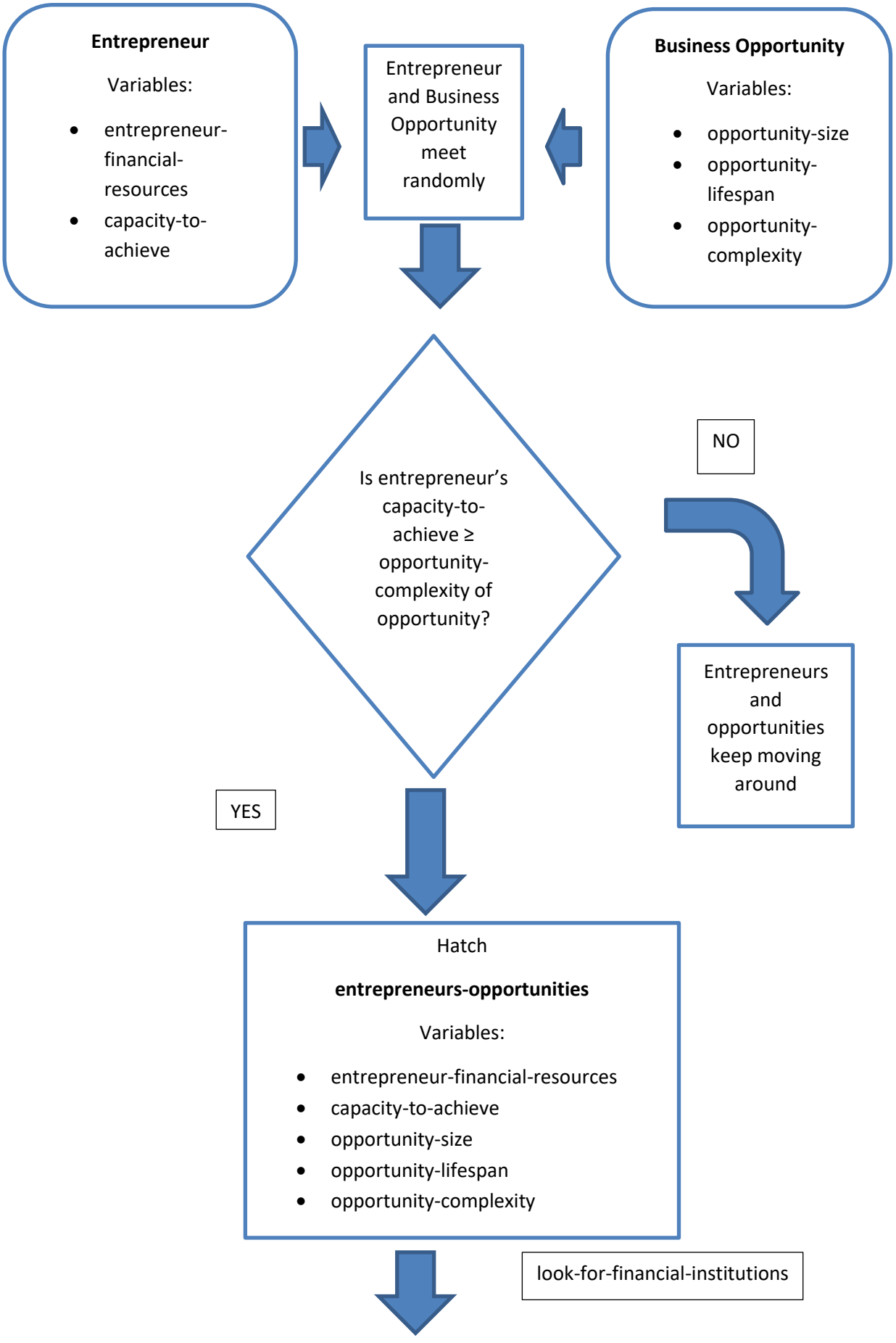
Currently, the model can be download from **Modeling Commons**, at:

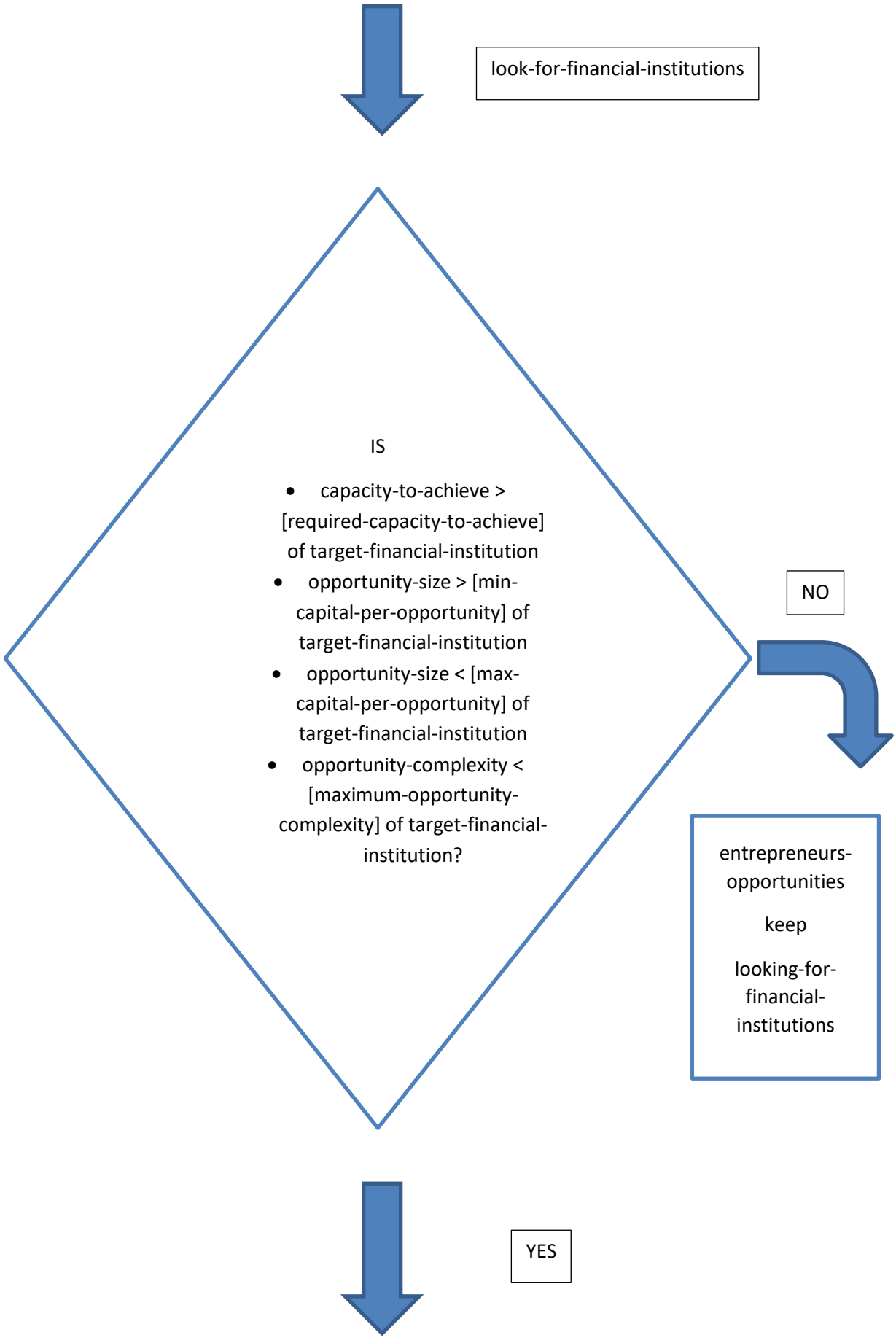
- Link: http://modelingcommons.org/browse/one_model/5715
- Access: To be provided. The repository requires the e-mails of the people who is going to access to this model ("share").
- [The model is not publicly available yet: please, keep the access and password safely.]

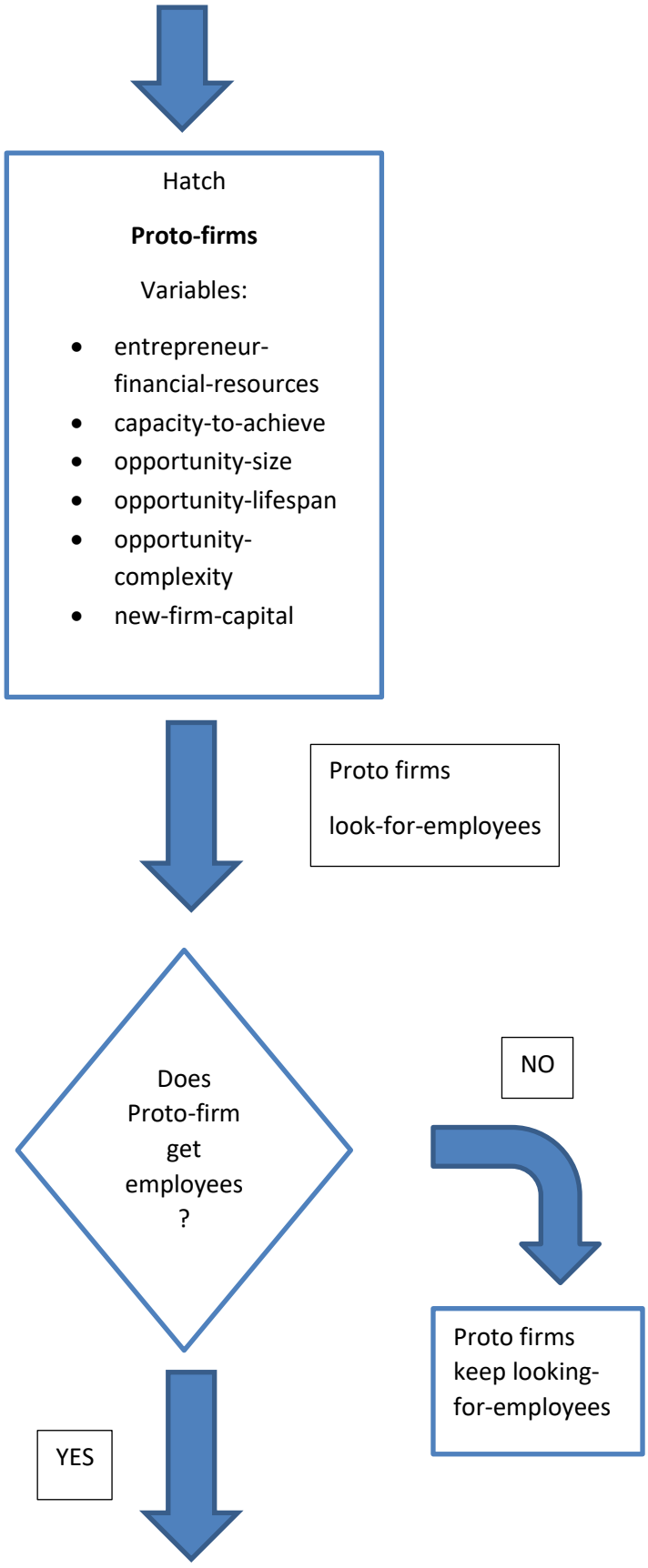
APPEARANCE OF THE GRAPHICAL INTERFACE

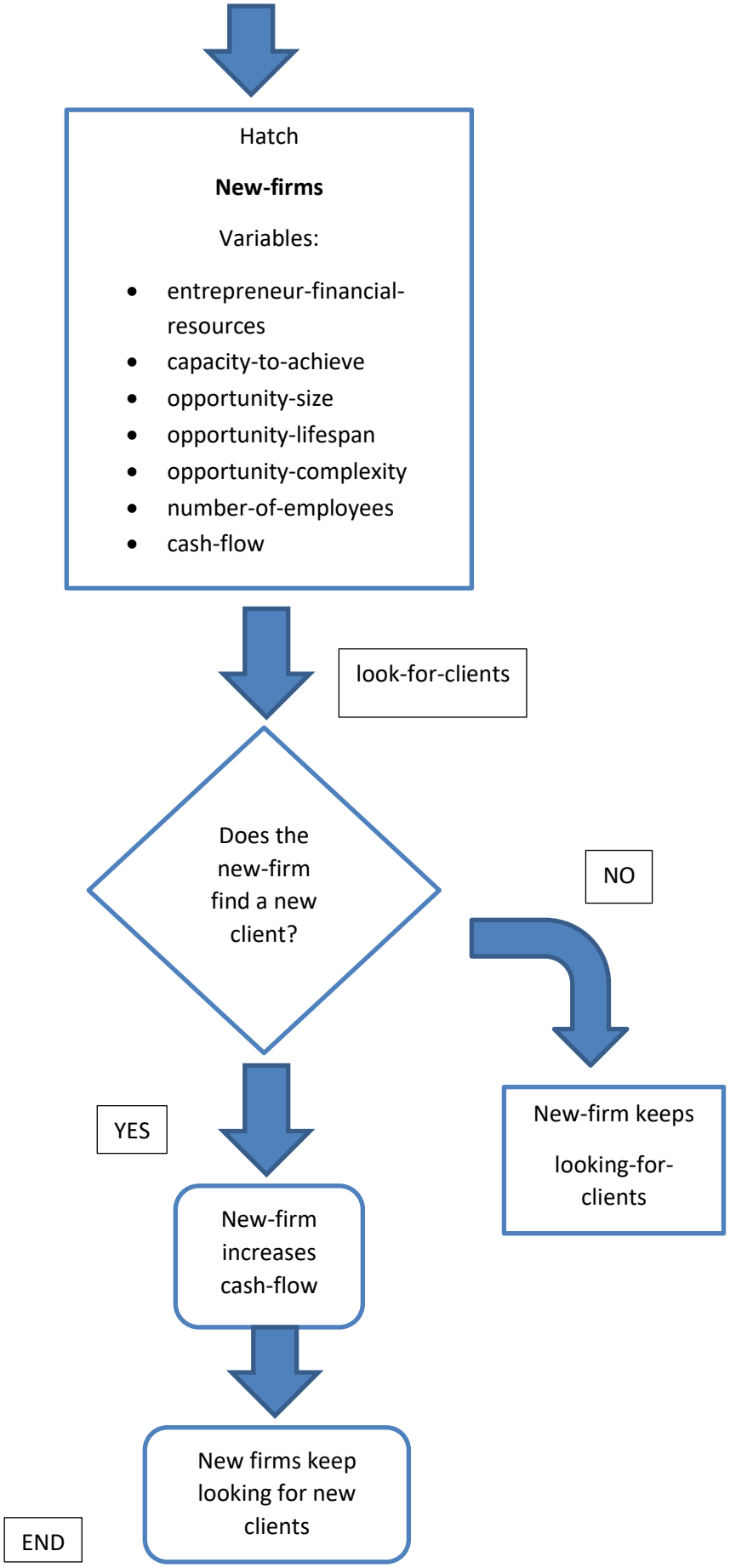


FLOW CHART


Entrepreneur
Variables:
entrepreneur-financial-resources
capacity-to-achieve
Entrepreneur and Business Opportunity meet randomly
Business Opportunity
Variables:
opportunity-size
opportunity-lifespan
opportunity-complexity
Is entrepreneur's capacity-to-achieve ≥ opportunity-complexity of opportunity?
NO
Entrepreneurs and opportunities keep moving around
YES
Hatch
entrepreneurs-opportunities
Variables:
entrepreneur-financial-resources
capacity-to-achieve
opportunity-size
opportunity-lifespan
opportunity-complexity
look-for-financial-institutions


look-for-financial-institutions
IS
capacity-to-achieve > [required-capacity-to-achieve] of target-financial-institution
opportunity-size > [min-capital-per-opportunity] of target-financial-institution
opportunity-size < [max-capital-per-opportunity] of target-financial-institution
opportunity-complexity < [maximum-opportunity-complexity] of target-financial-institution?
NO
entrepreneurs-opportunities
keep
looking-for-financial-institutions
YES


Hatch
Proto-firms
Variables:
entrepreneur-financial-resources
capacity-to-achieve
opportunity-size
opportunity-lifespan
opportunity-complexity
new-firm-capital
Proto firms look-for-employees
Does Proto-firm get employees ?
NO
Proto firms keep looking-for-employees
YES


Hatch
New-firms
Variables:
entrepreneur-financial-resources
capacity-to-achieve
opportunity-size
opportunity-lifespan
opportunity-complexity
number-of-employees
cash-flow
look-for-clients
Does the new-firm find a new client?
NO
New-firm keeps looking-for-clients
YES
New-firm increases cash-flow
New firms keep looking for new clients
END

7.2. TRACE DOCUMENT

This is a TRACE document ("TRAnsparent and Comprehensive model Evaludation") which provides supporting evidence of initial testing of our model presented in the PhD thesis titled:

"Exploring log-normal distributions in nascent entrepreneurship outcomes: International comparisons and agent-based modelling"

The rationale of this document follows:

> Schmolke A, Thorbek P, DeAngelis DL, Grimm V., 2010. Ecological modelling supporting environmental decision making: a strategy for the future. *Trends in Ecology and Evolution,* 25, pp. 479-486.

and uses the updated standard terminology and document structure in:

> Grimm V, Augusiak J, Focks A, Frank B, Gabsi F, Johnston ASA, Kułakowska K, Liu C, Martin BT, Meli M, Radchuk V, Schmolke A, Thorbek P, Railsback SF., 2014. Towards better modelling and decision support: documenting model development, testing, and analysis using TRACE. *Ecological Modelling,* vol. 280, pp. 129-139.

And

> Augusiak J, Van den Brink PJ, Grimm V., 2014. Merging validation and evaluation of ecological models to 'evaludation': a review of terminology and a practical approach. *Ecological Modelling,* vol. 280, pp. 117-128.

We have also kept the simple formatting of the author's template (including the letter font type and size) in order that TRACE documents produced by different authors keep the same structure and terminology - as in any standard format -.

Regarding the use of this protocol in the scientific community (especially in the EU funded research), see also:

http://cream-itn.eu/trace

http://cream-itn.eu/creamwp/wp-content/uploads/Trace-Guidance-11-03-04.pdf

The TRACE document, as a protocol, is self-contained: it is designed to be attached to the agent-based computer file as a complete explanation of the objectives and mechanisms of the model. Therefore, some repetitions of previous themes of this thesis have to appear.

Contents

PROBLEM FORMULATION

Relatively recent analyses of longitudinal nascent entrepreneurial panels have revealed the pervasiveness of the presence of heavy tailed distributions in their inputs and outputs (Crawford et al., 2015, Shim, 2016; Shim et al., 2017). In many datasets, lognormal distributions or power law distributions with an exponential cut-off can be plausible fit. However, the mechanisms that generate these heavy-tailed distribution patterns remain still poorly understood. Researchers have proposed a combination of multiplicative processes and/or preferential attachment to explain the results (Breig, Coblenz and Pelz, 2018).

This agent-based model is a research tool that has been designed to allow entrepreneurial researches to test their theories about nascent entrepreneurial processes and their heavy-tailed distribution patterns using the empirical datasets of

the current 14 different ongoing longitudinal panel projects worldwide, and to adapt, parametrize and calibrate their own models.

The generality of our baseline model enables future simulations with different parameters depending on the country/region under study. Our model is complex enough to integrate the diversity of parameters and conditions of the different countries in which the empirical longitudinal panel are implemented. It also allows easy changes in code to explore different assumptions. Our conceptual model is flexible permitting changes in conceptual framework, procedures, schedules, values ranges, and in the set-up state variables and behavior of the agents.

The baseline model – and further developments of it - will be openly available at the two main public agent-based model repositories to the entrepreneurship research community and nascent entrepreneurial stakeholders. The background material and code will be made available on permanent repositories such as the CoMSES Compu-tational Model Library maintained by the OpenABM consortium (http://www.openabm.org/models) and on the Modeling Commons (http://modelingcommons.org/), a Web-based collaboration system for NetLogo modelers.

MODEL DESCRIPTION

The model description follows the ODD (Overview, Design concepts, Details) protocol for describing individual-based models (Grimm et al., 2006; Grimm et al., 2010). The model was implemented in NetLogo 6.0.4 (Wilensky, 1999), a free software platform for implementing agent-based models. The NetLogo code will be made available on the permanent repositories CoMSES Computational Model Library, maintained by the OpenABM consortium (http://www.openabm.org/models) and on "Modeling Commons", a public space for online modeling in NetLogo, developed by the Center for Connected Learning and Computer-Based Modeling ("CCL") at Northwestern University.

- CoMSES Computational Model Library maintained by the OpenABM consortium: (http://www.openabm.org/models)
- Modeling Commons: (http://modelingcommons.org/)

ODD PROTOCOL OF "A NASCENT ENTREPRENEURIAL AGENT-BASED MODEL"

1. PURPOSE OF THIS MODEL

Crawford et al. published in the *Journal of Business Venturing* (Volume 30, Issue 5, September 2015, pages 696-713) a paper titled: "Power law distributions in entrepreneurship: Implications for theory and research". Their study analyzed three datasets in the United States, the Panel Study of Entrepreneurial Dynamics (PSED II), The Kauffman Firm Survey (KFS), the Inc. Magazine 5000 list, and one Australian data set (CAUSEE, The Comprehensive Australian Study of Entrepreneurial Emergence). They examined the distribution of key variables in nascent entrepreneurship, such as revenues, number of employees, number of owners, resources, etc., and they found that the majority of the variables showed a heavy-tailed distribution, specifically power law distributions, according to the distribution pitting techniques available at that time.

As part of the first section of this PhD research, I have also conducted an analogous study on the Swedish dataset of nascent entrepreneurs (by Mikael Samuelsson). Results show similar heavy-tailed distribution patterns and parameters to those founded in the United States and Australia datasets, although our statistical distribution pitting analysis showed that both power law and lognormal distributions can be reasonable fit for the different national datasets available to date.

One of the most recent attempts of the simulation of entrepreneurial outcomes distributions was initially developed by Shim (2016) using R software. He performed a simulation to determine if heavy-tailed distributions can be obtained through multiplicative processes in entrepreneurship. Shim (2016) was able to show that the distributions of the simulated outcomes were quite similar to

the empirical datasets and that lognormal models have better fit than other heavy-tailed distributions in most of the nascent venture early stages (activities) results. However, Shim (2016) suggested that more sophisticated agent-based modelling and simulations were needed, given that a simple random multiplicative process was not enough to explain the complexity of the empirical and simulated patterns.

Based on a bibliometric method and on the behavioural rules inferred from the entrepreneurship literature, Shim, Bliemel and Choi (2017) proposed a basic agent-based model that was able to reproduce the emergence of heavy-tailed distributions in nascent venture outcomes and that was consistent with the empirical datasets. Their model consists only in two agents ("entrepreneur" and "investor") and two objects ("opportunity" and "resources"), being the amount of resources modelled as state variables of entrepreneurs and investor. Breig, Coblenz and Pelz (2018) has recently proposed another agent-based model used as an illustrative example of statistical validation for the entrepreneurial variable "venture debt" with the empirical data extracted from the second Panel Study of Entrepreneurial Dynamic (PSED II).

However, in order to explore more complex phenomena in nascent entrepreneurship or to introduce other important elements of this nascent entrepreneurial process, a more complex agent-based model is required. Although complexity science researchers have identified several causal processes that yield heavy-tailed distributions in natural and social phenomena, the explanation for these distributions in nascent entrepreneurial processes requires further exploration (Breig, Coblenz and Pelz, 2018).

This model is designed to explore questions regarding the emergence of new ventures and their nascent entrepreneurial processes, and to identify the mechanisms that produce the emergence of heavy-tailed distributed outcomes ("patterns", Grimm, 2005) in nascent entrepreneurs' longitudinal data panels (PSED and similar empirical datasets). Although this model adopts most of the basic features and conceptual framework used in previous models (especially the

conceptual model of Gartner, 1985, and the roadmap proposed by Yang and Chandra, 2013), it introduces new levels of complexity in comparison to previous ones (Shim, 2016; Shim, Bliemel and Choi, 2017; Breig, Coblenz and Pelz, 2018).

There are new features, more internal state variables for agents, new forms of interactions among them, new rules of behavior, and types of agents, global environmental variables (Martinez, Yang and Aldrich, 2011), that allow the possibility of further research in relationship with the empirical data: calibration, parametrization, verification, etc. One of the purposes of this model is to expand previous "stylized fact" type of agent-based modeling - based on basic principles - to richer representation of real-world scenarios based on empirical datasets. A more complex model also allows deeper theory development from simulation (Davis et al., 2007). Our model starts with the discovery of the heavy tailed distribution patterns at the macro level – the "stylized fact" -, and it tries to simulate the underlying processes and behaviors of individual entrepreneurs at the micro level that produce that "stylized fact" (the pattern, the heavy tailed distribution) (Shim, Bliemel and Choi, 2017).

The conceptual model (inspired by Gartner, 1985; Yang and Chandra, 2013) is the following:

- There are two types of initial mobile agents in the Netlogo's world: entrepreneurs and opportunities. Both agents operate in a torus-like square grid.
- Entrepreneurs "search" opportunities through serendipitous discovery. In further developments of the model, entrepreneurs "sniff" opportunities (opportunities leave "tracks", like "pheromones" in biology).
- An entrepreneur encounters a business opportunity in the Netlogo's world. The theoretical framework is Shane and Venkataraman's (2000) opportunity "discovery-evaluation-exploitation".

- If their characteristics match, the entrepreneur tries to exploit the opportunity. They both become a dual entity “entrepreneur-opportunity” and this duet begins the start-up activities:

-
    - First, they look for financial institution to get money to implement the opportunity.
    - Second, the look for employees.
    - Third, the look for clients to increase their cash-flow.

- The run stops when only new firms remain in the “world”, and the rest of the mobile agents have “died”.

INITIAL MOBILE AGENTS

At the very beginning, the model has only two mobile agents: **Entrepreneurs** and **Business Opportunities.**

ENTREPRENEURS

Number of entrepreneurs:

Defined by a slider in the interface of the model "population-of-entrepreneurs" (for validation and calibration purposes, subsequent models will use real, empirical datasets (PSED, CAUSEE, etc.)).

Entrepreneurs´ state variables:

entrepreneur-financial-resources:

- The personal investment capital owned by the entrepreneur him/herself or his/her proxies (family, friends, etc.).
- Heuristically, in the first baseline model, it goes from 0 to 3,000,000 monetary units (originally 100,000). It follows the real empirical ranges described in PSED II, under the variable "individual investment" (Crawford et al, 2015, Table 1, p. 703).
- The amount is assigned to each entrepreneur randomly in the first baseline model.
- A preliminary sensitivity analysis showed that this amount is able to change the dynamics of the process greatly. Further analysis is required.
- A further development of the model will allocate entrepreneur's financial resources according to a lognormal distribution, instead of random distribution. To give each agent a number from a log-normal distribution,

we will use this code from Hamill and Gilbert (2016, documentation Chapter 7):

- set entrepreneur-financial-resources precision (e ^ random-normal mean standard-deviation) 0)

  Where function "precision *number places"* reports *number* rounded to *places* decimal places. Example:

  show precision 1.23456789 3

  => 1.235

  And, according to Crawford et al. (2015, Table 1, p. 703), variable "individual investment" has a mean = 23 and a standard-deviation = 110 (Curtin, 2012).

- For validation and calibration purposes, subsequent models will use real, empirical datasets (PSED, CAUSEE, etc.). The model has already incorporated the code to introduce the individual valued of this variable into the state variables of the agents (file name: "entrepreneur-financial-resources-EmpiricalData.txt") (see procedure "to setup-entrepreneurs").
- It could be also possible to generate a random value using different distributions (for example, a power law, as Crawford et al. (2015) suggested).

capacity-to-achieve:

- It represents the entrepreneur´s social and human capital, strong and weak tie networks (Granovetter, 1973; Gordon and Jack, 2010), "small world networks" (Watts and Strogats, 1998; Watts, 1999; Uzzi et al., 2007), acquaintance with investment capital, opportunity recognition capabilities, entrepreneur's education, previous experience in industry or venture founded, genetic factors (Nicolaou and Shane, 2009), etc.
- This property will affect the next step of contacting and matching with opportunities, venture capital institutions and banks.

- Numerally, it goes from 0 to 100, being 100 the highest capacity to achieve (a percentage scale). A high value increases the possibility of getting investments from others.
- The amount is assigned randomly to each entrepreneur in the first baseline model. Further developments of the model may assign this amount following a different distribution (lognormal or power law).
- For validation and calibration purposes, subsequent models may use real, empirical datasets (it may require the development of scales similar to those designed by Crawford et al. (2015, Appendix 1, "Construct, variables, and items", p. 710).

BUSINESS OPPORTUNITIES

The integration of the entity "business opportunities" into the model is one of the more challenging aspect of the development of this research tool because their elusive nature (Dimov, 2011). This agent requires further both theoretical development and practical implementation (state variables). Different options have been considered, such as, for example, to create an opportunity-generator "entity" in the model that would mimic industrial clusters, universities incubators, etc., and that would produce opportunities during the run using a distribution function (such as a random Poisson distribution). We have decided to keep things simple in the first baseline model which already has high level of complexity. In the baseline model, opportunities appear initially located physically and randomly in the world (in "clouds" shapes). Further refinement is needed.

Number of "Business opportunities":

Heuristically, it is defined by a slider with "number-of-opportunities" in that world.

Business Opportunities state variables:

opportunity-size:

- Investment capital (money) that is initially needed – individual entrepreneur's investment plus venture debt - to be successful in implementing the opportunity.
- Value: monetary units (depending on the country to be modelled, Euros, US dollars, Australian AUD, etc.).
- Heuristically, the amount is assigned randomly to each opportunity with a minimum of 100,000 monetary units and maximum of 5,100,000 approx. (originally, with a minimum of 100,000 monetary units and maximum 1,100,000). The range takes into consideration the empirical data of PSED II on the maximum value of the variable "Venture Debt" (Crawford et al., 2015, Table 1, p. 703).
- A preliminary sensitivity analysis showed that this amount is able to change the dynamics of the process greatly. Further analysis is required.

opportunity-lifespan:

- Time during which the opportunity is available, without being implemented.
- Units in times steps ("clicks"). After certain number of clicks, the opportunity is outdated and dies.
- The empirical panels (i.e. PSED) consider a span of 5 years as maximum. In the baseline model, each tick is a month (5 years * 12 months = 60 months).

opportunity-complexity:

- It is the counter-part of entrepreneur-own "capacity-to-achieve". A high "opportunity-complexity" value requires a high entrepreneur´s "capacity-to-achieve" in order to match. It reflects the need of many different resources (technological, financial, human, etc.) for implementing the opportunity.

For example, it is not the same to market a small plastic children toy than to market a new development for airplane wings.

- Numerally, it goes from 0 to 100, being 100 the highest "opportunity complexity" (a percentage scale).
- The amount is assigned randomly to each opportunity in the first baseline model.

Alternative model 1:

The model starts running with a defined numbers of entrepreneurs and opportunities (sliders) randomly located in the space and with both types of agents moving randomly around.

Alternative model 2 (not in this baseline version):

Entrepreneurs and Opportunities are generated from determined patches during the run following a determined probability function (like an "Opportunity generator").

STATIONARY AGENTS

Organizations (referred as "patches" in NetLogo). In the baseline model, there are three kinds of special patches.

FINANCIAL INSTITUTIONS: patches where entrepreneurs can obtain financial resources. They represent institutions such as banks, venture capital offices, investors, etc.

Number of financial institutions:

The number of financial institutions is defined by a slider in the interface.

They are randomly located in the world.

Financial institutions state variables:

financial-institution-resources:

- Amount of money ready to be invested. Total monetary units that the financial institution is able to lend.
- Heuristically, it goes from 100,000 to 10,100,000 monetary units approximately. It follows the real empirical ranges described in PSED II (see Curtis 2012, Codebook) (Crawford et al, 2015, Table 1, p. 703).
- It decreases every time the institution invests on a project, in the invested amount.
- Randomly assigned.

max-capital-per-opportunity:

- It sets the superior, maximum limit of investment in an opportunity (defined by "opportunity-size") of this concrete financial institution.
- Maximum 1,100,000 monetary units, approximately (minimum 100,000).
- Randomly assigned.

min-capital-per-opportunity:

- It sets the minimum limit of investment in an opportunity (defined by "opportunity-size").
- Minimum to invest: 50,000 monetary units (until 100,000). Heuristic (inverse calibration).

required-capacity-to-achieve:

- The financial institutions (investors, banks, venture capital, etc.) require a minimum of capacity-to-achieve in the entrepreneur, that is, his or her social and human capital, strong and weak networking, acquaintance with investment capital procedures, education, capacity of opportunity analysis, attitudes, knowledge of the sector, etc.
- Minimum 10 out of 100. Maximum 90 out of 100. Randomly assigned.

maximum-opportunity-complexity:

- Some investors may prefer big challenges, or the opposite.
- This is the maximum opportunity-complexity tolerated by the investors.
- Maximum 100 out of 100.
- Randomly assigned.

WORKFORCE AGENCIES: patches where employees can be hired. They represent employment offices (private or public), head-hunters, etc.

Number of workforce agencies:

The number of workforce agencies is defined by a slider in the interface.

They are randomly located in the world.

Workforce agencies state variables:

target-workforce-agency

- This version uses the solution of the Netlogo library example "Move towards Target Example" to resolve the problem related to the decision on which workforce agency should the new agent go, after being hatched.
- Randomly, a workforce agency is chosen by the new hatched agent to get employees.

CLIENTS: patches where the entities can obtain cash-flow.

Number of clients in the world:

The number of clients is defined by a slider in the interface.

They are randomly located in the world.

Clients state variables:

client-revenues:

- Amount of money that can be transferred to a new-firm from this client.
- Maximum 15,000 monetary units, minimum 5,000 monetary units, per commercial transaction (Heuristic amount, inverse parametrization).

Further developments of the model regarding organizations.

Two different organizations can be added:

- Opportunities generators: patches from where opportunities hatch. These patches represent universities, business schools, business incubators, industrial clusters, etc. They may generate opportunities following certain probability function (random Poisson, lognormal, etc.).
- Gathering patches: patches in which the probabilities of being linked to other agents or resources increase. They represent places such as business incubators, where entrepreneurs can meet other entrepreneurs, investors, opportunities, etc.

Geographically, the current conceptual framework of the model locates organizations distributed randomly in the plane – strictly from a topological point of view, in a torus - (XY patches, the plane of the "world", organizations with physical locations in the real world, "patches" in Netlogo language), such as banks, business incubators, universities, business schools, industrial clusters, manufacture plants of suppliers, clients, employment offices, workforce companies, head-hunters, etc. However, given the possibility of creation a three-dimensional world in Netlogo, organizations can also be in the space (XYZ patches, organizations that can be accessed through virtual links – internet, etc. -). The third dimension can be used to explore "weak" entrepreneurial networks, such as webpages of venture capital, crowdfunding websites, social networks of entrepreneurs, business networks (LinkedIn), etc.

SUBSEQUENT MOBILE AGENTS

ENTREPRENEUR-OPPORTUNITY

- If an entrepreneur encounters an opportunity and their variables match, a new agent is formed ("hatched"), the "ENTREPRENEUR-OPPORTUNITY" new entity.
- It gathers the properties of the parents' entrepreneur and the opportunity. When the new agent Entrepreneur-Opportunity is hatched, its parents die (the agent entrepreneur and the agent opportunity).

Entrepreneur-Opportunity state variables:

entrepreneur-financial-resources:

- This new agent has its parents' entrepreneur´s financial resources or what it is left, because the entrepreneur spends money in his/her wanderings.

capacity-to-achieve:

- The new agent has its parents' entrepreneur capacity-to-achieve.
- This property will determine the next step of contacting and matching with financial institutions (venture capital institutions and banks).
- The higher, the more possibilities of getting investments from others.
- Numerally, from 1 to 100, being 100 the higher capacity to achieve.

opportunity-size:

- It retains the variable value from the original parent's opportunity.

opportunity-lifespan:

- Only the ticks that are left from the parent's opportunity will pass onto this new agent.

opportunity-complexity:

- It retains the variable value from the original parent's opportunity.
- This property will affect the next step of contacting and matching with venture capital institutions and banks. Some investor may prefer big challenges or the opposite.
-

target-financial-institution

- Randomly, a financial institution is chosen to get financial resources.

PROTO-FIRM

- When an "entrepreneur-opportunity" meets and matches a financial-institution, there is a probability of becoming a proto-firm (being hatched) and of receiving the needed capital to implement the business opportunity (stochastic process, following Simon´s approach). A new agent (breed "proto-firm") is hatched from their

parents, the entrepreneur-opportunity and the contribution of the financial institution.

- The first baseline model does not introduce the probability function of becoming a proto-firm in order to avoid the increase of complexity at the beginning of the development of this entrepreneurial model.
- Proto-firms have capital from the financial institutions but they do not have workers yet, therefore, they are not fully operational and they cannot attend clients. Proto-firms will have to look for workers, first.

Proto-firms state variables:

entrepreneur-financial-resources:

- This agent has the entrepreneur´s financial resources or what it is left, because the parent entrepreneur-opportunity spends money in his/her wanderings.

opportunity-size:

- It retains the value from the parent's entrepreneur-opportunity.

opportunity-lifespan

- Only the time steps (ticks) that are left from its parent's entrepreneur-opportunity pass onto this new breed.

opportunity-complexity:

- It retains the variable value from the original parent's entrepreneur-opportunity.

new-firm-capital:

- The value of this variable is the sum of the entrepreneur-financial-resources of the parent's "entrepreneur-opportunity" plus "opportunity size".
- The amount of money "opportunity-size" (defined as the needed capital – money - to be successful implementing the opportunity) comes from the financial institution.
- This capital decreases as the new hatched agent is looking for employees.

target-workforce-agency

- Randomly, a workforce agency is chosen to get employees.

capacity-to-achieve:

- The new agent has its parent entrepreneur-opportunity's capacity-to-achieve.
- At this point, the value of this variable is not necessary, but it will be kept for tracking and research purposes.

NEW FIRM

- When a proto-firm gets employees it becomes a new-firm. It is hatched from the encounter between a proto-firm and a workforce agency, being the parent the proto-firm.
- New firms look for clients and get revenues.

New firms state variables:

new-firm-capital:

- It retains the variable value of the parent's proto-firm (It is the sum of the "entrepreneur-financial-resources" of the parent "entrepreneur-opportunity" plus "opportunity size").
- This capital decreases as the agent goes around looking for employees, and moving around searching for the initial client.

cash-flow

- This variable is made of the sum of the variables new-firm-capital (from the proto-firm parent) and client-revenues (in the encounter with clients).
- It decreases at every time step ("tick") due to the cost of searching for clients.
- It also decreases at every time step ("tick") due to the cost of employees´ salaries. The baseline model does not include this feature yet.
- If it is negative, the new firm is broke, and it dies.

number-of-employees.

- The initial number of employees has a relationship with the "new-firm-capital".
- The bigger the initial capital, the more the initial number of employees.
- Calculation: number-of-employees = 0.00001 * new-firm-capital. The concept “revenues per employee” (RPE) is an operating performance ratio, and it is recorded in the annual reports - or form 10-K in USA -. It indicates productivity levels and effective use of the firm’s resources. The range goes from smaller firms that average around $100,000 per employee versus almost $300,000 for a Fortune 500 company. For example, WalMart averages $170,000 revenue per employee; GE is around at $436,000 per employee; Microsoft averages $646,000 per employee; and the oil industry generates over $2 million per employee. This performance ratio can be obtained in standard business databases (in USA, for example, D&B Hoovers). The empirical longitudinal panels will provide the percentages for each country and new venture. This parameter is the inverse of this value: 1/(revenue/number of employees) = 1/ RPE (“revenues per employee”- RPE). It depends of the industry and of the “intensive in labor” nature of the firm (Microsoft versus Wallmart). Heuristically, we assume a “revenues per employee” average of a small company: around $100,000 per employee (1/100,000 = 0.00001).
- In further developments of the model, the numbers of employees should increase as the revenues increase.

target-client

- This version uses the solution of the Netlogo library example "Move towards Target Example" to resolve the problem of to which clients should the new-firm go, after becoming a firm.
- In this baseline model, the client is going to be chosen randomly. So, it can repeat the same client again and again (depending on a random function, and, therefore, there is low probability of this event).

capacity-to-achieve:

- The new firm has its parent's capacity-to-achieve.
- At this point, the value of this variable is not necessary, but it will be kept for tracking and research purposes.

Spatial context:

The initial baseline model will consist in a "city", or a "region" or a "country", depending of the level of geographical detail provided by the empirical micro-data from the panels (PSED, CAUSEE, etc.). The size of the grid cells would be calculated accordingly (for example, 1 grid cell = 10 km^2, so on and so forth).

The base model uses Sweden as example for further parametrization and calibration. Sweden has a populated area of around 200,000 km2 (40% of its land; the North has a very low population). Our grid is 500 km^2 x 500 km^2, with 50 patches per square side with value of 10 km^2 per patch. ($\sqrt{200{,}000}$ km^2 ~ 500 km^2).

In 2D -coordinates X and Y, the plane where the "world" is- some grid cells are agents: some patches of land represent the geographical locations of physical entities involved in the nascent entrepreneurial dynamics such as banks, business incubators, universities, business schools, industrial clusters, manufacture plants of suppliers, major clients, employment offices, workforce companies, head-hunters, etc.

In the current version of the code of the model, only three types of patch agents are represented by the grid cells: financial institutions, workforce agencies and clients. Overlap of roles occurs: a grid cell is a static agent with its own variables, but also it functions as a location in the "world".

Temporal scale:

The temporal step value (the "tick") depends on the system under study. For example, in US PSED, one temporal step ("tick") represents one day, and simulations run for 5 years (60 months, 21900 days). Empirical longitudinal panel datasets also include the date in which each activity was performed by the nascent entrepreneur. The panel empirical data sets (PSED, CAUSEE, etc.) will serve as method of calibration for the model in future developments. Each empirical agent data recorded by the longitudinal panel can be introduce into the model with code for importing files similar to the one indicated in the code section relative to the setup of entrepreneur's state variables.

Current baseline model - based on PSED - consider a span of 5 years in which each tick is a month (5 years * 12 months = 60 months).

Environmental variables

The baseline model has global variables that influence the "world" and that affect to all agents. These global variables represent aspects described in reports such as those by the Global Entrepreneurship Monitor (GEM), "Doing Business" (World Bank Group), etc., related to the business environment and entrepreneurial enhancers. The variables associated to business environment included in the panels (PSED, etc.) can also be taken into account (Martinez, Yang and Aldrich, 2011).

The global variables can be set with sliders in NetLogo interface in each run.

Currently, two environmental global variables have been coded in the initial interface of the baseline model:

search-cost:

- Money spent in each of the entrepreneurial actions: discovery and development of opportunities, transactions, preparation of business plans, travelling, networking, etc. ""Search" is costly and may influence the success/failure of entrepreneurs" (Yang and Chandra, 2013, p. 214).
- Each tick has a defined temporal value depending on the longitudinal panel data set (a day, a month, etc.). The slider defines the cost of actions during that day/month/etc. It includes the entrepreneur´s salary (or cost of living).

social-dynamism:

- The slider defines the number of steps in the two-dimensional grid made by the agents in the "word" in every tick: dynamic business environment versus slow environment.

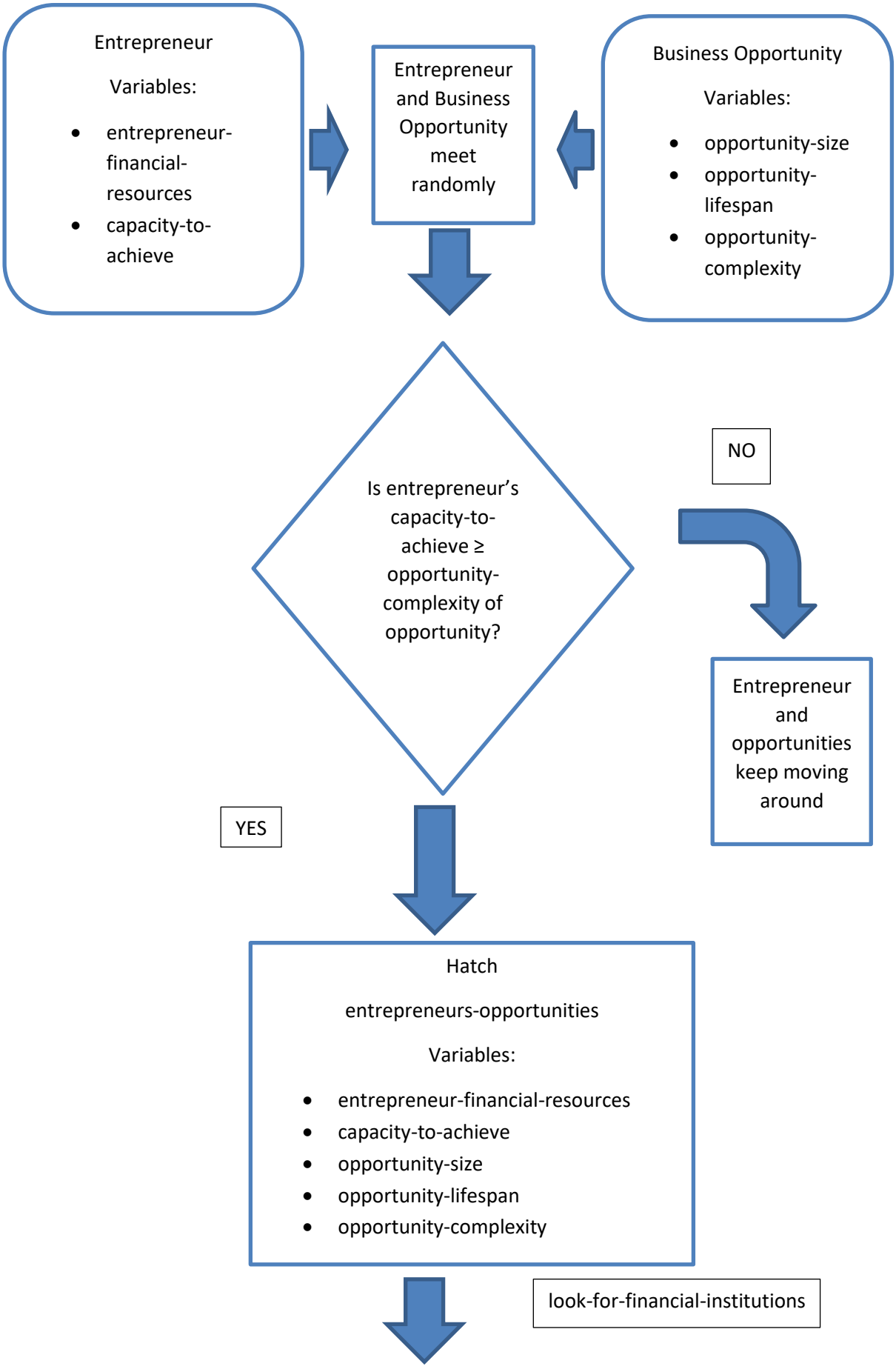
- "In most social and business phenomena, space, or distance plays a crucial role which will determine the emergence of an event or not and how/why it occurs" (Yang and Chandra, 2013, p. 213).

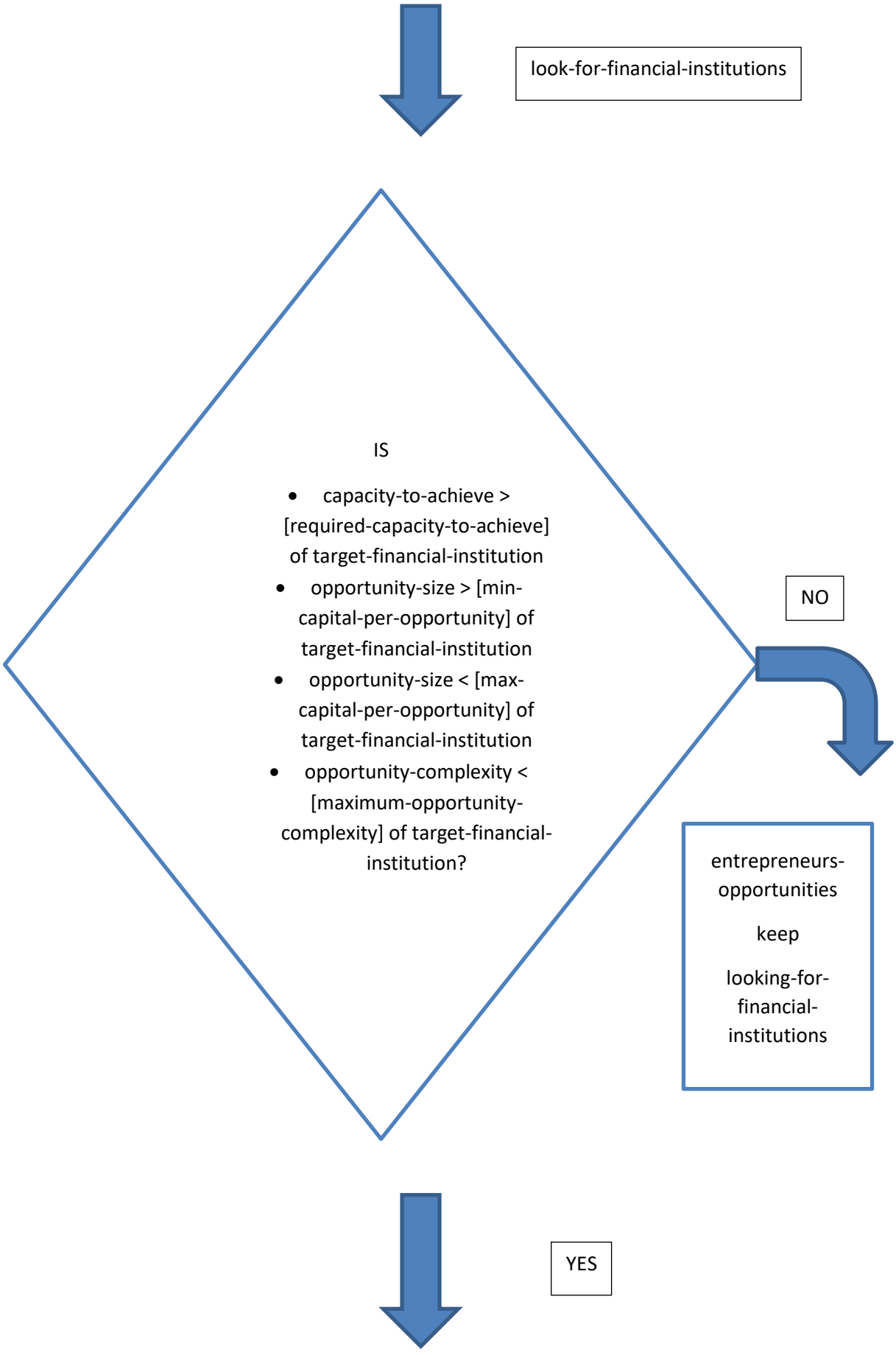
These other potential environmental variables can also be introduced in the model, in future versions:

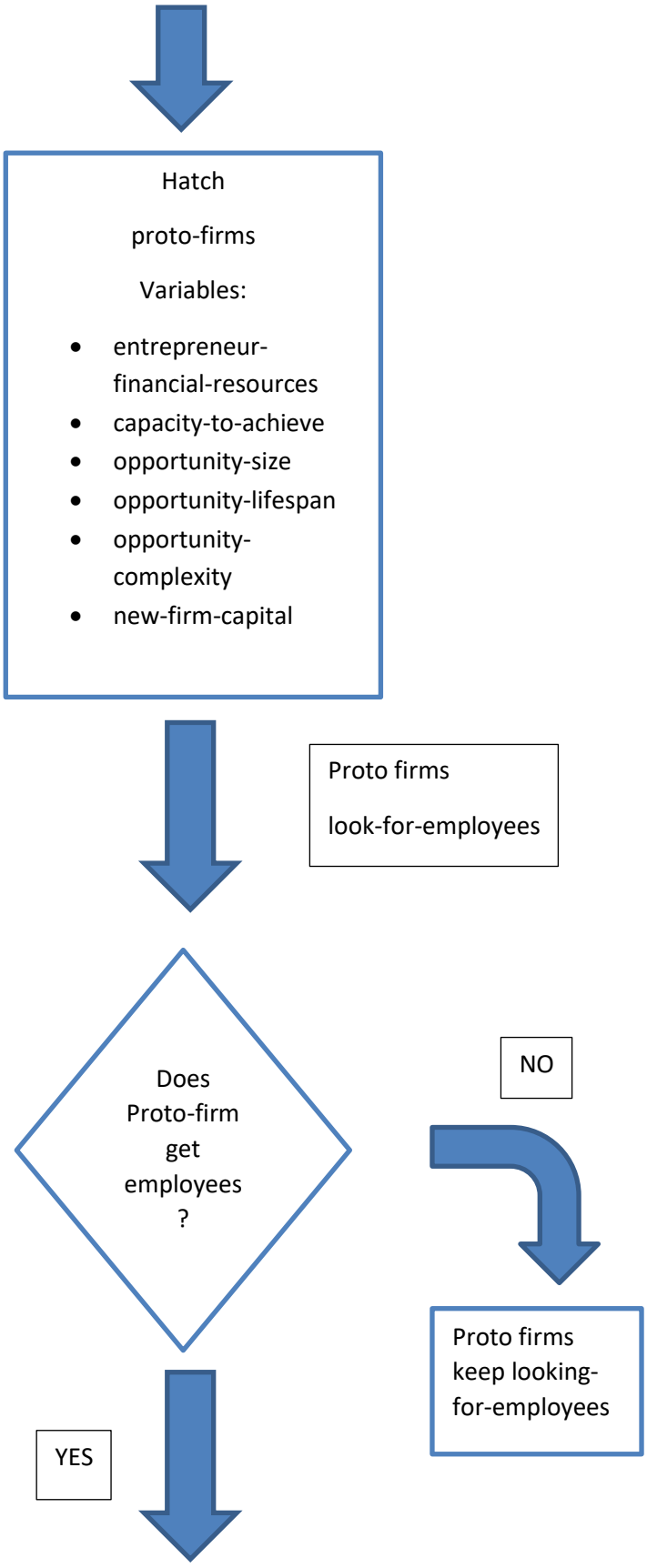
- Entrepreneurial Public Policies: this global variable encompasses aspects such as regulations, taxes, registration conditions, administrative constrain, etc. A favourable entrepreneurial public policy would increase the possibility of exploiting opportunities, becoming a firm, etc.
- Entrepreneurial Business Environment: this global variable encompasses aspects such as the degree of entrepreneurial activity, numbers of industrial clusters, social approval of entrepreneurial activity, active R&D institutions, availability of financial resources etc. A favourable Entrepreneurial Business Environment increases the number of opportunities that hatch, the number of entrepreneurs in the simulated "world", etc.
- Tendency of entrepreneurs to cooperate and to become teams. It increases the probability of becoming a team when one or more entrepreneurs meet which, in turn, it increases the probability that a team of entrepreneurs can exploit bigger opportunities, etc.
- Capital growth: the capital in the financial institutions grows every year a certain percentage. To be added as a slider. Also it can be coded as a process "regrow-resources".

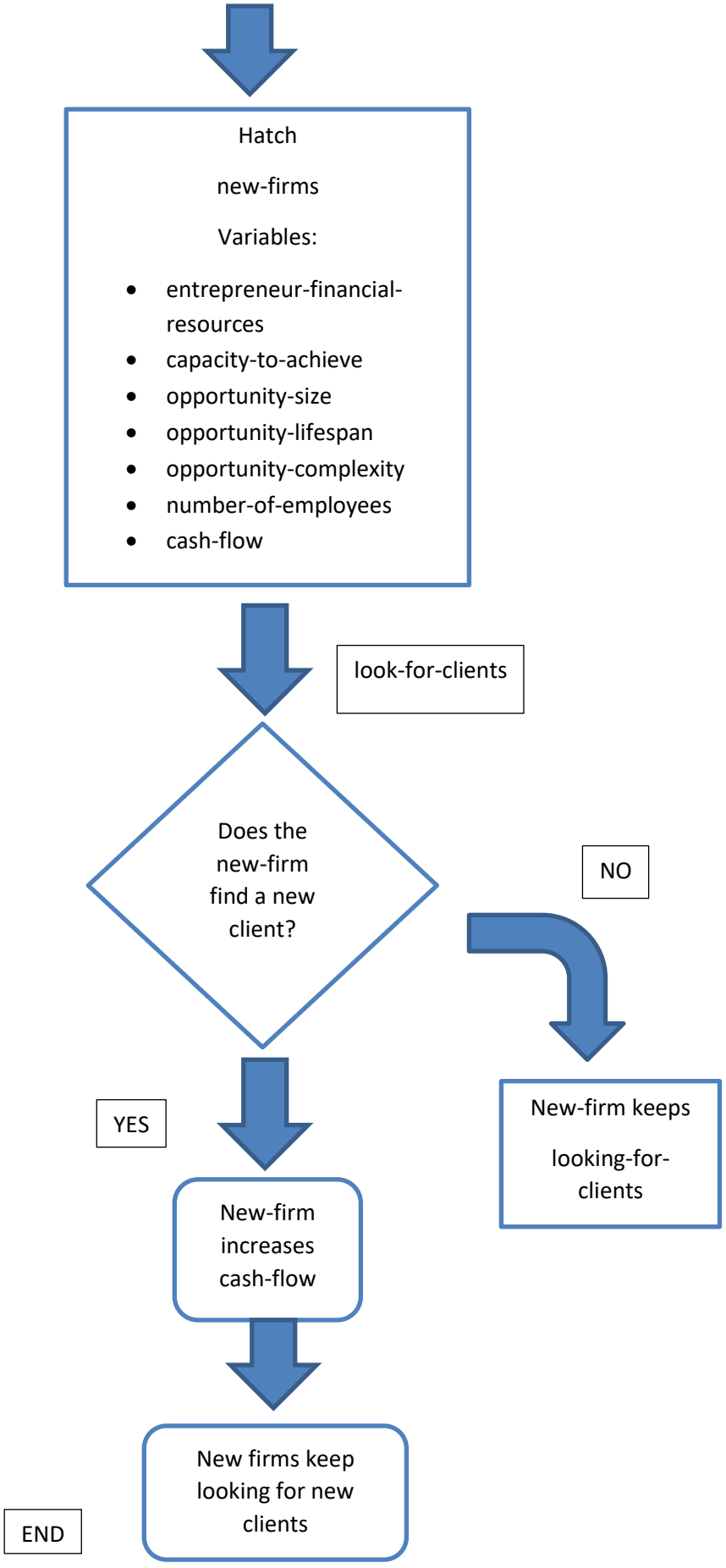
2. PROCESS OVERVIEW AND SCHEDULING

The following flow chart shows the summary of the process and scheduling:


Entrepreneur
Variables:
entrepreneur-financial-resources
capacity-to-achieve
Entrepreneur and Business Opportunity meet randomly
Business Opportunity
Variables:
opportunity-size
opportunity-lifespan
opportunity-complexity
Is entrepreneur's capacity-to-achieve ≥ opportunity-complexity of opportunity?
NO
Entrepreneur and opportunities keep moving around
YES
Hatch
entrepreneurs-opportunities
Variables:
entrepreneur-financial-resources
capacity-to-achieve
opportunity-size
opportunity-lifespan
opportunity-complexity
look-for-financial-institutions


look-for-financial-institutions
IS
capacity-to-achieve > [required-capacity-to-achieve] of target-financial-institution
opportunity-size > [min-capital-per-opportunity] of target-financial-institution
opportunity-size < [max-capital-per-opportunity] of target-financial-institution
opportunity-complexity < [maximum-opportunity-complexity] of target-financial-institution?
NO
entrepreneurs-opportunities
keep
looking-for-financial-institutions
YES


Hatch
proto-firms
Variables:
entrepreneur-financial-resources
capacity-to-achieve
opportunity-size
opportunity-lifespan
opportunity-complexity
new-firm-capital
Proto firms
look-for-employees
Does Proto-firm get employees ?
NO
Proto firms keep looking-for-employees
YES


Hatch
new-firms
Variables:
entrepreneur-financial-resources
capacity-to-achieve
opportunity-size
opportunity-lifespan
opportunity-complexity
number-of-employees
cash-flow
look-for-clients
Does the new-firm find a new client?
NO
New-firm keeps looking-for-clients
YES
New-firm increases cash-flow
New firms keep looking for new clients
END

The following section describes what entity does, and in what order, using **pseudo-code** to describe the schedule of the process. A state variable is immediately assigned a new value as soon as that value is calculated by a process; therefore, the model follows an **asynchronous updating**.

Entrepreneurs processes:

Entrepreneurs meet opportunities by chance, randomly, by wandering around the "world" (human behavioral ecology approach). [This process implies a complex “foraging” theory approach to entrepreneurship and it has yet to be justified. It would introduce another complementary biological perspective on entrepreneurship - human behavioral ecology - in addition to genetics, physiology and neuroscience (Shane, 2009; Nofal, Nicolaou, Symeonidou and Shane, 2018). Also related with to organizational ecology approach: Hannan and Freeman, 1977; Freeman et al., 1983; Freeman and Hannan, 1983).]

- Wiggle: first turn a little bit randomly.
- move-entrepreneurs: then, step forward. This movement implies spending resources (as "search-cost").
- check-if-broke: check to see if entrepreneurs has already spent their own "entrepreneur-financial-resources", because the step forward implies consumption of resources, as it is defined in the interface slider "search-cost".
- get-opportunity: entrepreneurs that look for opportunities. If the characteristics of both match, they may become new breed: “entrepreneur-opportunity”.
- [meet-other-entrepreneurs: in the first baseline model, this process has not been coded yet. It implies the formation of team of entrepreneurs. For future model developments].

Opportunities processes:

They move around randomly. In the initial baseline model, they exist at the beginning of the run and they are located randomly in the "world". Future model development will simulate "opportunities generators" at locations such as universities, business cluster, etc.

- Wiggle.
- move-opportunity.
- check-if-outdated: every opportunity has randomly assigned a different lifespan of a number of ticks. Opportunities get old and outdated.

"entrepreneurs-opportunities" breed processes.

- look-for-financial-institutions: entrepreneurs-opportunities look for money to implement the project.
- check-if-outdated: opportunities are out of date after a certain amounts of ticks.
- [check-for-becoming-a-new-firm: it may happen that the entrepreneurs´ owns financial-resources are enough to become a new firm. The conditions will be: entrepreneur-financial-resources >= opportunity-size. Not coded yet for simplification purposes].

Proto-firms processes:

- look-for-employees: proto-firms look for a workforces agency to get employees.
- check-capital-level: the search for employees consumes resources, the defined “search costs”. If new-firm-capital is below a certain threshold, the proto-firm is brought to an end. If the “new-firm-capital” variable is less or equal to zero, the proto-firm dies.

New-firms processes:

- look-for-clients: New-firm entities look for clients to increase cash-flow.
- check-for-liquidation: if new-firm-revenues are below a certain threshold, involving variables such as "cash-flow" and "new-firm-capital", the new-firm is dissolved ("bankruptcy").

my-update-plot:

- Update the different plots of the interface. In the first baseline model, the update after the “tick” is done directly in the interface for practical reasons. Future models will update from the code section (good coding practices, Railsback and Grimm, 2011).

Stop the run

- The run stop when the agents have reached the new-firm status or the intermediate agents have already disappeared.
- If the number of entrepreneurs is zero and the number of entrepreneurs-opportunities is zero, and the number of proto-firms is zero, the run stops.

4. DESIGN CONCEPTS

Basic Principles.

The main hypothesis of the model is that the individual behavior of the agents and the multiplicative nature of their interactions in nascent entrepreneurship processes may explain the emergence of heavy tailed distributions (power laws or lognormal) in the observed empirical data. The theoretical framework for agent traits, from which system dynamics emerge, have been described in the different subsections of this document.

The "rules of engagement" of the agents follow Gartner's theoretical framework (1985) for describing new venture creation. This theory is particularly suitable for this research that tries to develop a model of venture emergence (it is described above). This model also assumes a "foraging theory", taking a behavioral ecology approach regarding the entrepreneurs encountering business opportunities: these two agents are "wanderers" in the model's "world".

The basis of the conceptual framework of the model is the following:

- An entrepreneur encounters a business opportunity in the Netlogo's world.
- If their characteristics match, they begin the start-up activities (the temporal order of these activity can easily be changed in the code if theory testing is desired):
    - First, they look for financial institution to get money to implement the opportunity.
    - Second, the look for employees.
    - Third, the look for clients to increase their cash-flow.
- At every start-up activity, the state variables of agents have to match in order to have a successful result (money, employees). The run stops when only "new firms" remain in the "world", and the rest of the mobile agents have "died".

The empirical datasets of the different entrepreneurial longitudinal panels across the world will be used to calibrate de model.

Emergence.

The objective of the model is to reproduce the emergence of heavy tailed distributions observed in the experimental datasets of different longitudinal panels of nascent entrepreneurship. The emergence of these heavy tailed distributions is not obvious or trivial. The model will experiment with different mechanisms and parameters that generate these types of distributions to identify the ones that better match the empirical results.

The baseline model uses stochasticity in the assignation of variable values. In the first baseline model, a random generator is used (Netlogo primitive called

"random"). However, the model results show that, even with a random variable value generation, heavy-tailed distributions **emerge** after undergoing the nascent entrepreneurial processes modelled: power laws, log-normal and even Weibull distributions can be obtained varying the parameters. Surprising, the experiments conducted were not able to reproduce Gaussian distributions easily. The normality tests normally failed when analyzing the results of the model runs. Future models should experiment with the implementation of a random lognormal generator. Stochastic processes may yield heavy-tailed distributions under certain conditions.

Adaptation.

The agents of the model have different adaptive traits depending on the entrepreneurial phase in which they are:

- An entrepreneur seeks a suitable opportunity.
- An entity [Entrepreneur-Opportunity] seeks a financial resource.
- A Proto-firm seeks a workforce agency.
- A new-firm looks for clients to get revenues.

The hierarchy of start-up activities to be implemented by the entities in each period is based on the prevalence of the activities that has been determined by empirical data of the panels (for example, the PSED start-up activities prevalence in Reynolds, 2017b).

In the current coded baseline model, these searches of the agents (for opportunities, financial institutions, workforce agency, or clients) are purely stochastic. Further developments in the models should experiment introducing

some selection, looking for the most suitable match, and observing how this may change the distributions of the state variables and outputs.

Objectives.

The objective of the entities is to become a firm, with clients, employees, and a positive cash flow. Entities prioritize certain start-up actions depending on the step they are in the process of emerging venture (see "Adaptation" section above).

Learning.

In the initial baseline model, changes in adaptive behavior as a consequence of experience (learning) has not been considered for sake of simplicity. However, among entrepreneurs' state variables, entrepreneurial experience is included as part of the human capital, increasing the frequency of gestation activities over time.

Prediction.

The purpose of this entrepreneurial model is to explain the heavy tailed distributions of entrepreneurial variables (revenues, number of employees, etc.) observed in empirical datasets. The baseline agent-based model uses randomness as the mechanism of making the decisions of agents. For example, a proto-firm selects randomly one of the workforce agencies to get employees; an agent entrepreneur-opportunity also picks up one of the financial institutions randomly.

However, although this randomness in the processes may be founded in empirical evidence in entrepreneurship (Coad 2009; Coad, 2013; Frankish et al., 2013, p. 77; Lotti et al., 2009), it also may impose an assumption that may lead to the predicted heavy tailed distribution. Different mechanisms may be playing an important role in the emergence of heavy tailed distributions that can be hidden by this random, stochastic approach, such as the Yule process (preferential attachment), or the critical phenomena and the associated concept of self-organized criticality (Newman, 2005, p. 348-349).

Sensing.

In this model, agents can "sense" the state variables of the other agents and of the patches - where organizations are located - once they are in them. An entrepreneur may sense if an opportunity is suitable for her/him or not when he/she encounters it. For example, an entrepreneur's capacity-to-achieve has to be bigger or equal to the state variable "opportunity-complexity" of the target-opportunity to be able to implement the sub-model "to get-opportunity". Likewise, a financial institution may "sense" and reject an "Entrepreneur-Opportunity" entity that does not fit with its investment portfolio criteria.

The baseline model makes the agents to choose a target randomly (a financial institution, a workforce agent, a client) to keep things as simple as possible. A further development of the model may use the model principle of diffusion, similarly to the one use in the NetLogo Ants model (Wilensky, 1997). Using the Netlogo primitive "diffuse", it is possible for the institutional patches to send information about their own resources around. This modeling technique would allow the introduction of the "money-smell" and to give the capacity of the agents to "sniff" (sense) the state variables of the organizations that they are most interested. These trails generates by the Netlogo primitive "diffuse" also provide

information to the agents regarding of the source of these trails. For example, a trail to a venture capital patch informs the agent of the amount of financial resource potentially available if this trail is followed.

Interaction.

The baseline model has direct interactions, in which agents encounter others, and, if some conditions are met, they are able to hatch a new agent ("breed", in Netlogo terminology). In further developments of the model, an indirect interaction will be implemented through competition, for the revenues of the clients, or for the financial resources of the financial institutions (venture capital).

Stochasticity.

Stochasticity has been the major source of variability in the Simon's tradition on firms' size distribution. The basic mechanism for generating power laws, for example, has been proportional random growth (Gabaix, 2009; Gabaix, 2014; see subsection above in this document).

The baseline model uses the Netlogo primitive "random" that produces what is called in computer programming a "pseudo-random" number, through a deterministic process (by the generator known as the *Mersenne Twister*). In scientific modeling, pseudo-random numbers are better because if the model start with the same random "seed" is possible to get the same results every time, and, therefore, to develop experiments that can be reproduced by other researchers. The code to implement the random seed is already in the coding section, before the set-up procedure. This feature makes the model run reproducible (see Replication section above).

An improvement of the initial model would also use random processes to cause events or behaviors to occur with a specified frequency such as the hatching of the opportunities and their characteristics, the frequency of an entity proto-firm of becoming a firm, etc. [still to be coded].

Besides the uniformly distributed random integers generated by the primitive "random", NetLogo also offers several other random distributions such as random-normal, random-poisson, or it is able to generate other random distributions through code (see code for random-lognormal above). Future developments of this model should include these other types of distributions to analysis the impact in the model (as suggested in the previous section).

Collectives.

In this model, collectives are merely a definition of the types of agents (breeds), characterized by their own state variables. Organizations such as clients, financial institutions or workforce agencies have their own state variables assigned randomly in the set-up.

Emergent collectives out of the individuals' behavior are not expected.

Further developments of the model will include the possibility to create entrepreneurs' teams (not coded yet, for simplification purposes).

Observation.

The main outcomes collected from the model for analysis are:

- Start-up survival (time to be born and time to die) (not coded yet).
- Number of established firms after the runs.
- Number of employees.
- Cash flows.

The model is inspired by several empirical longitudinal panels developed in different countries. These panels selected a significant sample of entrepreneurs in a country, and follow them through a number of years registering their entrepreneurial activity. The model tries to capture and understand the main output of those empirical longitudinal datasets. The model has already implemented the code to export data in CSV (Comma Separated Value) files for statistical study, in this case, for distribution pitting with R (for example with package '**Dpit'** version 1.0: Joo, Aguinis and Bradley, 2017 or "**fitdistriplus"**), or other distribution testing packages (such as '**goft'** version 1.3.4: Gonzalez-Estrada, & Villasenor-Alva, 2017). Detailed data on every step of the model can also be obtained through the Netlogo *BehaviourSpace* software tool.

There are two possible options for the implementation of this observational principle: 1) all the output data are used, or 2) only certain data sample is used, replicating the methodology of the longitudinal panels. Further calibration tests are required to decide the better approach.

5. INITIALIZATION

There will be two main versions of the model is the initial state (at time t = 0 of a simulation run:

1) An artificial "world", as an initial test for the baseline model.
    a. Initially, there will be only two initial mobile agents, entrepreneurs and opportunities.
    b. The variables of the different agents will be set stochastically from a defined range (see table with parameters ranges).
    c. The initial conditions are also established by sliders, with the number of agents and global variables values (such as "search-cost", etc.).

2) A "world" based on the one of the empirical datasets provided by the longitudinal panels (parametrization - calibration). The empirical datasets of PSED cover five years. At the initial state of the model world (time t=0 of a simulation run), the numbers of entities and the values of their state variables will be set based on the panels empirical data at year 0. The model has already a mute code for initialization data introduction in the set-up procedure ("file-read" primitive) [not implemented in the baseline version]

6. INPUT DATA

The initial baseline model does not use input data to represent time-varying processes (time series) such as the environmental variables changes related to business environment (changes in legal frameworks, financial crisis, financial bubbles, etc.). However, further developments of the model would introduce inputs related to the environmental, global variables relative to business environment, public policies, regulations, etc. such as the ones described in the Global

Entrepreneurship Monitor Reports (GEM), “Doing Business Report Series” (World Bank Group), etc.

7. SUBMODELS

This section described in detail the submodels that represent the processes listed in ‘Process overview and scheduling’. The description includes Netlogo’s code (computer language) or pseudo-code to make possible a replication.

Please notice that, in Netlogo, the semicolon (;) mutes code and text. It is used to make programming comments, explanations, or to mute code that can be activated later. The semicolons in this section mean that the text after this syntax symbol is muted code or comments/explanations.

Submodel “moving around the world”: to wiggle.

to wiggle: entrepreneur and opportunity procedure.

- The initial process of finding the right opportunity by the entrepreneur is random, stochastic, just "good luck".
- This is an assumption based on some results from bibliography (Coad, 2009; Coad, 2013; Frankish et al., 2013, p. 77; Lotti et al., 2009). It should be taken with a "grain of salt". Further theoretical development is needed to justify it.
- Code:
    - right random 90 ;; turn randomly right
    - left random 90 ;; turn randomly left
    - if not can-move? social-dynamism [right 180]: It does not leave the "world". To avoid a violation of topology.

Submodel "to move entrepreneurs"

to move-entrepreneurs: entrepreneur procedure.

- Step forward the number of steps defined by a slider at the interface called "forward social-dynamism". It was defined as the number of steps in every tick: dynamic business environment *versus* slow business environment. However, preliminary extreme tests proved that there is something problematic with this global variable "social-dynamism". The problem may be related to the topology and the way agent's search each other in the model. It affects the performance of the model greatly, but not as it was initially thought. This variable remains in the baseline model for further testing.
- This movement implies spending resources (as "search-cost"). Set entrepreneur-financial-resources (entrepreneur-financial-resources - search-cost). Moving has a financial cost.

Submodel "check-if-broke"

to check-if-broke: entrepreneur procedure.

- If entrepreneurs-financial-resources are too low, he or she gets out of the game.
- Code:

```
if (entrepreneur-financial-resources <= search-cost) [die]
```

Submodel "to get an opportunity".

to get-opportunity: entrepreneur procedure (based on code from "Wolf Sheep Predation" in Netlogo model library (Wilensky, 1997):

- If an entrepreneur meets an opportunity, and this opportunity matches the required conditions, a new breed is hatched (an "entrepreneur-opportunity" entity), with the properties of its parents. Afterwards, the parent entrepreneur dies, as well as the opportunity that was found.
- Code:

  let target-opportunity one-of opportunities-here ;; procedure from the entrepreneur agent perspective.

  if target-opportunity != nobody ;; there is an opportunity in this patch with me (entrepreneur agent perspective).

  [

  if (entrepreneur-financial-resources >= search-cost) and (capacity-to-achieve >= [opportunity-complexity] of target-opportunity)

  [

  hatch-entrepreneurs-opportunities 1. This is a new breed. [Further development: Can we introduce a probability function such as hatch entrepreneur-opportunity random-poisson or random-normal? How different the model would behave? To be explored.]

The variables of the new breed entrepreneur-opportunity are made of a combination of the opportunity and entrepreneur´s variables.

The new breed variables are calculated in the following form:

Variables values coming from the entrepreneur:

- set entrepreneur-financial-resources (entrepreneur-financial-resources - search-cost) ;; to accept an opportunity has a cost (time to analyze it, decisions, etc.).
- set capacity-to-achieve (capacity-to-achieve)
  - This property will affect the next step of contacting and matching with venture capital institutions and banks.
  - The higher, the more possibilities of getting investments from institutions.

Variables values coming from the opportunity:

- set opportunity-size ([opportunity-size] of target-opportunity): it retains this property from the original target-opportunity.
- set opportunity-lifespan ([opportunity-lifespan] of target-opportunity): only the ticks that are left pass onto this new breed.
- set opportunity-complexity ([opportunity-complexity] of target-opportunity):
  - This property will affect the next step of contacting and matching with venture capital institutions and banks.
  - Some investor may prefer big challenges or the opposite.

New aspects of the new breed “entrepreneur-opportunity” are added such as size and where to go next through random “target” approach (primitive “one-of”):

Code:

- set size 2 (easier to see in the interface).
- set target-financial-institution one-of financial-institutions: set the financial institution target randomly to which it is going forward next.

- face target-financial-institution: point in the direction of the target, to go forward in the next step.

After taking its variables to become the new breed, the opportunity dies. The parent entrepreneur also dies.

Submodel "move-opportunity"

to move-opportunity: opportunity procedure: step forward.

- This movement decreases opportunity-lifespan. Opportunity-lifespan is measured in "ticks" units.
- forward social-dynamism ;; it defines the number of steps in every tick: dynamic business environment versus slow business environment. It is defined in a slider at the interface (See comments above).
- set opportunity-lifespan (opportunity-lifespan - social-dynamism): to exist implies a decrease of opportunity-lifespan until it get old and outdated. "Social-dynamism ticks" are subtracted in each tick in order to reflect that in very dynamic societies opportunities get older faster. Further study is required to analysis the impact of "social-dynamism" global variable in the whole model.

Submodel "check-if-outdated"

to check-if-outdated ;; opportunity procedure.

- Opportunities have an opportunity-lifespan: it is the time for the opportunity to remain available, without being implemented. Business opportunities get old, and eventually, they die (they are removed from the "world").

- Code:

  if (opportunity-lifespan <= social-dynamism), the opportunity dies (it is outdated).

Submodel "look-for-financial-institutions"

to look-for-financial-institutions: "entrepreneurs-opportunities" breed procedure.

This breed goes to financial institutions to get money.

The financial institutions set conditions to accept an "entrepreneur-opportunity" breed. These are the conditions to get the money:

(capacity-to-achieve > [required-capacity-to-achieve] of target-financial-institution),

and (opportunity-size > [min-capital-per-opportunity] of target-financial-institution),

and (opportunity-size < [max-capital-per-opportunity] of target-financial-institution),

and (opportunity-complexity < [maximum-opportunity-complexity] of target-financial-institution)) .

When an entrepreneur-opportunity meets and matches with a financial-institution, they may become a proto-firm (hatch-proto-firms 1), but it has not workers/employees yet (not fully operational). [Further development: Can we introduce a probability function such as hatch-proto-firms random-poisson or random-normal? How different the model would behave? To be explored.]

The new proto-firm states variables comes from the parent‘s “entrepreneur-opportunity”:

- set entrepreneur-financial-resources (entrepreneur-financial-resources): this agent has the entrepreneur´s financial resources or what it is left, because the entrepreneur spends money in her/his wanderings.
- set capacity-to-achieve (capacity-to-achieve): The higher, the more possibilities of getting investments.
- set opportunity-size (opportunity-size): it retains this property from the original opportunity.
- set opportunity-lifespan (opportunity-lifespan): only the ticks that are left pass onto this new breed proto-firm.
- set opportunity-complexity (opportunity-complexity).
- set new-firm-capital (entrepreneur-financial-resources + opportunity-size): it is the sum of the entrepreneur-financial-resources of the parent "entrepreneur-opportunity" plus the "opportunity size" amount. The “opportunity size” amount is provided by the financial institution, if the “entrepreneur-opportunity” state variables fulfill the profile of the financial institution portfolio.

A workforce agency is then randomly targeted to which it is going forward next. It points to the direction of the target, to go forward in the next step.

The resources of the financial institution decrease in the amount invested:

Code:

- set financial-institution-resources ([financial-institution-resources] of target-financial-institution - ([opportunity-size] of entrepreneur-opportunity))

The "entrepreneur-opportunity" entity is removed (it dies) in order to let it become a new breed "proto-firm".

If the "entrepreneur-opportunity" breed does not match the conditions of the financial institution, it targets another financial institution randomly (as long as it has resources to go around – search-cost -).

Further developments of submodel "look-for-financial-institutions": once the new-firm reach certain amount of cash-flow, it can return to financial institutions for more money (subsequent rounds of investing) and to the workforce agencies to get more employees [not implemented yet in the baseline model].

Submodel "look-for-employees"

to look-for-employees: proto-firms procedure.

When the proto-firm meets a workforce agency gets employees and it becomes a functional new-firm (hatch-new-firms 1); it can also be done with a probability function such as random-poisson or random normal (To be explored to understand the impact in the model).

Code of how to define the new variables of the new breed "new-firms":

- set new-firm-capital (new-firm-capital - search-cost): the costs of the transaction of getting the employees from the agency.
- set number-of-employees (0.00001 * new-firm-capital): The concept "revenues per employee" (RPE) is an operating performance ratio, and it is recorded in the annual reports - or form 10-K in USA -. It indicates productivity levels and effective use of the firm's resources. The range goes from smaller firms that average around $100,000 per employee versus

almost $300,000 for a Fortune 500 company. For example, WalMart averages $170,000 revenue per employee; GE is around at $436,000 per employee; Microsoft averages $646,000 per employee; and the oil industry generates over $2 million per employee. This performance ratio can be obtained in standard business databases (in USA, for example, D&B Hoovers). The empirical longitudinal panels will provide the percentages for each country and new venture.

- This parameter is the inverse of this value: 1/(revenue/number of employees) = 1/ RPE ("revenues per employee"- RPE). It depends of the industry and of the "intensive in labor" nature of the firm (Microsoft versus Wallmart). Heuristically, we assume a "revenues per employee" average of a small company: around $100,000 per employee (1/100,000 = 0.00001).
- set cash-flow (new-firm-capital)

The new firm sets the next client randomly to which it is going forward next ("target").

The parent proto-firm dies and it transfers its variable values to the new firm.

Further developments of this submodel ""look-for-employees": once the new-firm has reached a certain amount of cash-flow, it can return to the workforce agencies to get more employees.

Submodel "look-for-clients"

to look-for-client: new-firms procedure.

If a new-firm encounters a client, it gets client-revenues and the revenues are added to the cash-flow of the new-firm:

- set cash-flow (cash-flow + [client-revenues] of target-client - search-cost)

Afterwards, look (target) for another client (set target-client one-of clients). It is done randomly (Netlogo primitive “one-of” targets a new client randomly).

- set cash-flow (cash-flow - search-cost): there is a cost for searching new clients.

Submodel “check-capital-level”

to check-capital-level: proto-firm procedure. It checks if the amount of new firm capital is enough to look for clients. If not, it dies.

- if new-firm-capital <= 0 [die]

Submodel “check-for-liquidation”

to check-for-liquidation ;; new-firm procedure.

If cash-flow is negative, and this amount, in absolute value, is bigger than the capital of the firm, then, the new firm is dissolved ("bankruptcy").

Code:

if ((cash-flow <= 0) and (abs cash-flow >= new-firm-capital))

or new-firm-capital <= 0 [die]

Submodel "to R analysis"

The R-analysis button located in the interface of the model generates two output files:

1) cash-flow of new-firms at the end of the run.
2) number-of-employees at the end of the run.

The end of the run is defined when in the model there are only "new-firms" agents, and the entrepreneurs, "entrepreneurs-opportunities", and "proto-firms" have died.

Code:

- if (count entrepreneurs = 0) and (count entrepreneurs-opportunities = 0) and (count proto-firms = 0) [stop]

The model generates these two CVS files, "distributioncashflow.cvs" and "distributionofemployees.cvs" that can be easily imported into R for further statistical testing and analysis. Please notice that these two files are saved in the same folder where the Netlogo model is located in your system.

Submodel "R extension in Netlogo"

The processing of the data can also be done sending the data directly to R, using the R extension of Netlogo (Thiele & Grimm, 2010). Code can be found in this repository:

https://github.com/NetLogo/NetLogo).

Currently, the code for this extension is muted (using ;) because the implementation of the R extension in Netlogo requires further additional configuration depending on the operating system (Linux, Mac OS, Windows). In the muted coding, we have initially integrated the package Dpit and the "goft tests" (described above). It has also be coded some very useful visualization of the histograms and density function of the distribution (such as the Filled Kernel Density, in the R package "stats") relevant for further statistical analysis and comparisons (now muted in the code; they can be implanted just deleting the semicolon (;)).

DATA EVALUATION

Reminder - Relevant evaluation concepts in agent-based modelling addressed in this TRACE element:

- **Parameters** are the constants in the Netlogo's primitives, equations and algorithms that are used to represent the processes in an agent-based model.
- **Parameterization** is the task of selecting values for the parameters of a model to relate it to real system as much as possible (Railsback and Grimm, 2012, p. 255).
    - "Direct parameterization" is when parameter values are obtained directly from the literature or experts.
    - "Inverse parameterization", is when we define parameter values inversely by calibrating the model to reflect the real, empirical distribution, in this case, the heavy-tailed distributions (Grimm et al., 2014, p.4).

This baseline nascent entrepreneurial model is designed as a research tool, in which the parameters can be modified to tailor and to calibrate the model with the corresponding empirical longitudinal dataset at study. Most of the current parameters of the baseline model have been taken directly from the recorded empirical dataset of USA PSED or Sweden PSED. Therefore, although most of the parameters are direct, that is, taken from the empirical datasets, this baseline model is not fully calibrated yet: some parameter have been left open or flexible to adapt them to other datasets. The table below defines the ranges, values, units and references of the different parameters on our baseline model inspired on real data -but not fully parametrized -.

Some parameters are heuristic: they are reasonable assumptions out of experience.

Other heuristic values come from industry reports (for example, the value of the operating performance ratio “revenues per employee” (RPE)).

Some parameters and ranges that are very difficult to know in the real scenario have been assigned also heuristically by inverse calibration, trying to simulate the heavy-tailed pattern distributions of the real datasets (‘pattern-oriented modelling’; Grimm et al., 2005; Grimm and Railsback, 2012). The reliability of the parameters depends on the data gathering quality of each of the entrepreneurial longitudinal panel. The follow-up of hundreds of entrepreneurs during several years is a Titanic task. Many times, datasets are incomplete and/or not coherent.

The major potential sources of uncertainty in the model parameters correspond to the following elements:

- The real, empirical datasets themselves: they are full of statistical noise; incomplete, missing data, wrong/incoherent data, pitfalls of the interviewing process (operational mistakes, lies, data gathering mistakes, etc.).

- number-of-opportunities. To know the real number of business opportunities in a real context (a city, a country) is impossible, given the own nature of this concept (they are “elusive”: Dimov, 2011).

- Number of clients in the real world. The entrepreneurs have activities in different sectors of the economy. Some look for industrial clients, other for consumers, others for both. This baseline version of the model does not classify types of industry and sectors, for sake of simplicity. Further developments of the model can include those (with an "input file" coding – see example in the code) specifying the characteristics of each client in the Netlogo's world.

State variables of the financial institutions are still uncertain in this baseline model, such as

- financial-institution-resources
- max-capital-per-opportunity
- min-capital-per-opportunity
- maximum-opportunity-complexity
- required-capacity-to-achieve

Financial institutions can be very heterogeneous: it is not the same a venture capital specialized in software or technological start-ups, that the local branch of a commercial bank giving a loan to open a new butchery in town. On the other hand, the criteria for investment may change along time depending on resources or peculiarities of the investors/business angels. However, many of these variables of the financial institutions can be inferred doing some calculations of the variables measured in the longitudinal panel datasets. For example, the PSED includes several variables that ask about venture debt, bank loans, investors, etc. (Curtis 2012, Codebook). With some empirical data processing, the uncertainty of these variables can be substantially reduced.

Spatial context. This baseline model proposes a two-dimensional geographical grid (torus). It may work for a new small high street business. But

with the irruption of technology, this framework may be obsolete. A team of entrepreneurs in Bordeaux (France) may get investors located in Amsterdam (NL) thanks to internet and cheap flights or train tickets.

On the other hand, the grid reproduces a country such as Sweden or Australia, in which the population is concentrate in small portions of the territory. In Sweden most of the population is in the South (barely 40% of the territory; the North is almost unpopulated) and in Australia, the population in concentrated in only 27% of the country (the rest is also mostly unpopulated). The demographic distributions of other countries may be difficult to simulate regarding the spatial framework (for example, USA).

Temporal scale. Current baseline model is based on PSED temporal framework: it considers a span of 5 years in which each tick is a month (5 years * 12 months = 60 months). However, the Australian CAUSEE and the Sweden PSED use different time spans and different time intervals to interview the entrepreneurs' sample.

TABLE 5 - KEY PARAMETERS, VALUES AND REFERENCES – BASELINE: "A NASCENT ENTREPRENEURIAL AGENT-BASED MODEL"

Agent	Parameters State Variables	Values – Range-Distributions	Units	References	Further developments
Initial set-up	population-of-entrepreneurs	Number set by a slider in the interface	Entrepreneur	For calibration purposes: It depends of the empirical dataset at study (US PSED, Australian CAUSEE, Swedish PSED, etc.)	For calibration purposes: It will depend of the empirical dataset at study (currently there are 14 ongoing longitudinal panels similar to US PSED, Australian CAUSEE and Swedish PSED.
Initial set-up	number-of-opportunities	Number set by a slider in the interface	Opportunity	For calibration purposes: It depends of the empirical dataset at study (US PSED, Australian CAUSEE, Swedish PSED, etc.)	For calibration purposes: It will depend of the empirical dataset at study (currently there are 14 ongoing longitudinal panels similar to US PSED, Australian CAUSEE and Swedish PSED.
Initial set-up	number-financial-institutions	Number set by a slider in the interface	Financial institution	For calibration purposes: It depends of the empirical dataset at study (US PSED, Australian CAUSEE, Swedish PSED, etc.)	For calibration purposes: It will depend of the empirical dataset at study (currently there are 14 ongoing longitudinal panels similar to US PSED, Australian CAUSEE and Swedish PSED.
Initial set-up	number-of-workforce-agencies	Number set by a slider in the interface	Workforce Agency	For calibration purposes: It depends of the empirical dataset at study (US PSED, Australian CAUSEE, Swedish PSED, etc.)	For calibration purposes: It will depend of the empirical dataset at study (currently there are 14

					ongoing longitudinal panels similar to US PSED, Australian CAUSEE and Swedish PSED.
Initial set-up	number-of-clients	Number set by a slider in the interface	Client	For calibration purposes: It depends of the empirical dataset at study (US PSED, Australian CAUSEE, Swedish PSED, etc.)	For calibration purposes: It will depend of the empirical dataset at study (currently there are 14 ongoing longitudinal panels similar to US PSED, Australian CAUSEE and Swedish PSED.
Entrepreneur	entrepreneur-financial-resources	Random assignment from 0 – 100,000 (Here 100,000, by inverse calibration. Standard deviation real PSED = 110,000)	Monetary unit	It follows the real empirical ranges, for example those described in PSED II, under the variable "individual investment" (Crawford et al, 2015, Table 1, p. 703). University of Michigan (2018). For calibration purposes: It depends of the real data set at study (CAUSEE, Sweden PSED, etc).	Currently it is random. Other assignment distribution may be used.
Entrepreneur	capacity-to-achieve	Random assignment from 0% - 100%	Percentage	For validation and calibration purposes, subsequent models may use real, empirical datasets (it may require the development of scales similar to those designed by Crawford et al. (2015, Appendix 1, "Construct, variables, and items", p. 710).	The amount is assigned randomly to each entrepreneur in the first baseline model. Further developments of the model may assign this amount following a different distribution (log-normal or power law).
Opportunity	opportunity-size	Randomly assigned: minimum of 100,000 monetary units and maximum 1,100,000 aprox.	Monetary units	The amount is assigned randomly to each opportunity with a minimum of 100,000 monetary units and maximum of 5,100,000 approx (heuristically, currently, with a minimum of 100,000 monetary units and maximum 1,100,000 because of inverse calibration). The range takes into consideration the empirical data of	

		(inverse calibration)		PSED II on the maximum value of the variable "Venture Debt" (Crawford et al., 2015, Table 1, p. 703; University of Michigan, 2018).	
Opportunity	opportunity-lifespan	Randomly assigned: From 12 to 60 months.	Month	For calibration purposes: It depends of the empirical dataset at study (US PSED, Australian CAUSEE, Swedish PSED, etc.) and their interviews schedules. For example, the empirical panel USA PSED consider a span of 5 years as maximum. In the baseline model, each tick is a month (5 years * 12 months = 60 months).	
Opportunity	opportunity-complexity	Randomly assigned: From 0% - 100%	Percentage	For validation and calibration purposes, subsequent models may use real, empirical datasets (it may require the development of scales similar to those designed by Crawford et al. (2015, Appendix 1, "Construct, variables, and items", p. 710).	
Financial Institution	financial-institution-resources	Randomly assigned: from 100,000 to 10,100,000 (by inverse calibration)	Monetary unit	Heuristically, it goes from 100,000 to 10,100,000 monetary units approximately. For validation and calibration purposes, subsequent models may use real, empirical ranges described in PSED II (Curtis 2012, Codebook) (Crawford et al, 2015, Table 1, p. 703).	
Financial Institution	max-capital-per-opportunity	Maximum 1,100,000 monetary units, (minimum 100,000). Randomly assigned.	Monetary unit	Heuristic. (Inverse calibration)	
Financial Institution	min-capital-per-opportunity	Minimum to invest: 50,000 monetary units (until 100,000).	Monetary unit	Heuristic (Inverse calibration)	
Financial Institution	required-capacity-to-achieve	Minimum 10 out of 100. Maximum 90 out of 100. Randomly assigned.	Percentage	For validation and calibration purposes, subsequent models may use real, empirical datasets (it may require the development of scales similar to those designed by Crawford et al. (2015, Appendix 1, "Construct, variables, and items", p. 710).	
Financial	maximum-	Maximum 100 out of	Percentage	For validation and calibration purposes, subsequent	

Institution	opportunity-complexity	100. Randomly, assigned.		models may use real, empirical datasets (it may require the development of scales similar to those designed by Crawford et al. (2015, Appendix 1, "Construct, variables, and items", p. 710).	
Workforce agency	number-of-employees	0.00001 * new-firm-capital	Employee	The concept "revenues per employee" (RPE) is an operating performance ratio, and it is recorded in the annual reports -or form 10-K in USA-. It indicates productivity levels and effective use of the firm's resources. The range goes from smaller firms that average around $100,000 per employee versus almost $300,000 for a Fortune 500 company. For example, WalMart averages $170,000 revenue per employee; GE is around at $436,000 per employee; Microsoft averages $646,000 per employee; and the oil industry generates over $2 million. This performance ratio can be obtained in standard business databases (in USA, for example, D&B Hoovers).The empirical longitudinal panels will provide the percentages for each country and new ventures. This parameter is the inverse of this value: 1/(revenue/number of employees) = 1/ RPE ("revenues per employee"- RPE). It depends of the industry and of the "intensive in labor" nature of the firm (Microsoft versus Wallmart). Heuristically, we assume a "revenues per employee" average of a small company: around $100,000 per employee (1/100,000 = 0.00001).	
Client	client-revenues	Maximum 15,000; Minimum 5,000 monetary units per commercial transaction.	Monetary unit	Heuristic in the baseline model. For calibration purposes: It depends of the empirical dataset at study (US PSED, Australian CAUSEE, Swedish PSED, etc.)	
Global	social-dynamism	Number set by a slider in the interface	Step in the grid	Baseline: heuristic.	
Global	Search-costs	Number set by a	Monetary	For calibration purposes: It depends of the empirical	

		slider in the interface	unit	dataset at study (US PSED, Australian CAUSEE, Swedish PSED, etc.) The baseline model uses Swedish statistics. Average salary in Sweden: 3,000 GBP/month Average monthly cost of living in Sweden: 2,000 GBP/month Source: Swedish National Statistics: https://sweden.se/society/meet-the-average-anderssons/	
Global	Spatial context: two-dimensional grid (torus)	Baseline (example Sweden): 500 km^2 x 500 km^2, with 50 patches per square side with value of 10 km^2 per patch.	Km^2	For calibration purposes, it depends on the country characteristics. For example, Sweden has a populated area of around 200,000 km^2 (40% of its land; the North has a very low population). Our grid is 500 km^2 x 500 km^2, with 50 patches per square side with value of 10 km2 per patch. ($\sqrt{200,000}$ km^2 ~ 500 km^2).	
Global	Temporal scale	Current baseline model -based on PSED- consider a span of 5 years in which each tick is a month (5 years * 12 months = 60 months).	month	For calibration purposes: It depends of the empirical dataset at study (US PSED, Australian CAUSEE, Swedish PSED, etc.) and their interviews schedules. For example, the empirical panel USA PSED consider a span of 5 years as maximum. In the baseline model, each tick is a month (5 years * 12 months = 60 months).	

CONCEPTUAL MODEL EVALUATION

The conceptual model (based by Gartner, 1985; Yang and Chandra, 2013) is the following:

- There are two types of initial mobile agents in the Netlogo's world: entrepreneurs and opportunities. Both agents operate in a torus-like square grid.
- Entrepreneurs "search" opportunities through serendipitous discovery (random wandering throughout the world). In further developments of the model, entrepreneurs "sniff" opportunities –opportunities leave "tracks", like "pheromones" in biology. This is Gartner's "The entrepreneur locates a business opportunity" (1985).
- An entrepreneur encounters a business opportunity in the Netlogo's world. The theoretical framework is based on Shane and Venkataraman's (2000) opportunity "discovery-evaluation-exploitation".
- If their characteristics match - Shane and Venkataraman's (2000) "evaluation" -, the entrepreneur tries to exploit the opportunity. They both become a dual entity "entrepreneur-opportunity" and this duet begins the start-up activities:
    - First, they look for financial institution to get money to implement the opportunity (This is Gartner's "The entrepreneur accumulates resources" (1985)).
    - Second, the look for employees (These are Gartner's "The entrepreneur builds an organization " and "The entrepreneur produces the product" (1985))
    - Third, the look for clients to increase their cash-flow. (This is Gatner's "The entrepreneur markets products and services" (1985)).

- The run stops when only new firms remain in the “world”, and the rest of the mobile agents have “died”.

The flow chart and process overview and schedule is shown in the section 2.3 above.

Previous entrepreneurial agent model attempts already included two agents (“entrepreneurs” and “investors”) and two objects (“opportunities” and “resources”) (Shim, Bliemel and Choi, 2017). New agents and processes have been introduced introducing new levels of complexity in order to be able to simulate the nascent entrepreneurial processes in the different empirical longitudinal datasets.

The deep underlying theoretical framework of this PhD research is based in Behavioral Ecology (Aldrich, 2011; Davies, Krebs and West, 2012; Roundy, Bradshaw and Brockman, 2018). The encounter of the entrepreneur and the opportunity is a kind of complex “human foraging”.

The baseline model has been tested according to the guidelines suggested by Railsback and Grimm (2012) (new ed. 2019), Wilensky and Rand (2015), and Augusiak et al. (2014).

There are several pieces of defensive programming included throughout the model coding in order to avoid run-time errors or any other programming-related malfunction.

Specifically, the baseline model has been tested for:

- Typographical errors.
- Syntax error.
- Run-time errors.

  Notice that the plots in the interface with the histograms of the value of variables cash-flow and number of employees have some defensive coding to avoid run-time errors. The code is inside the plots themselves. To access it, “click” and “edit” on the plots. For example:

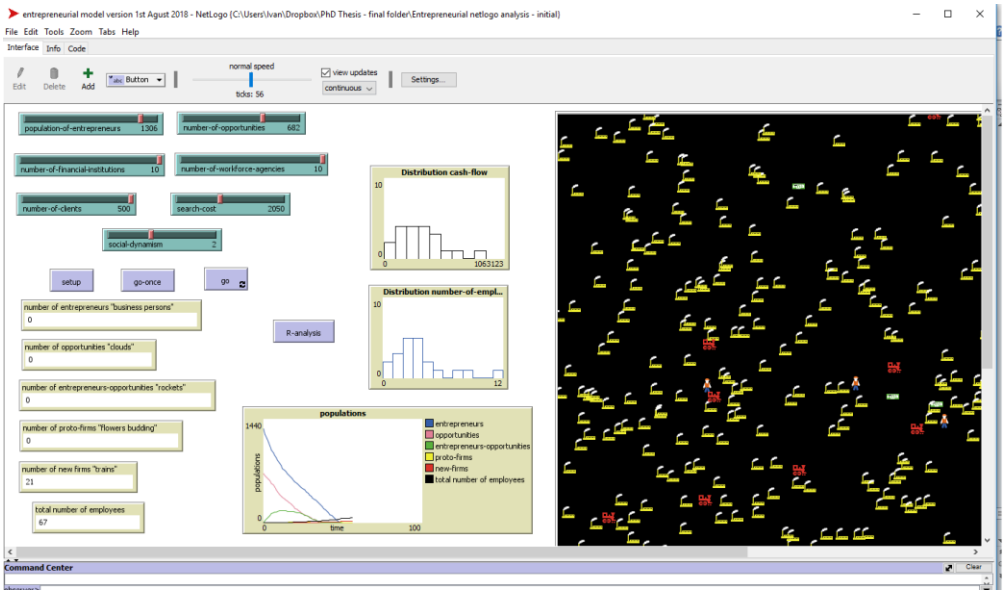
```
if not any? new-firms [stop]
let max-number-of-employees max [number-of-employees] of new-firms
plot-pen-reset  ;; erase what we plotted before
set-plot-x-range 0 (int (max-number-of-employees + 1))
histogram [number-of-employees] of new-firms
```

- Logic errors.

- Formulation errors. We have detected negative cash-flows in some runs. Further checks are required.
- Stress Test or “extreme” testing: running the model with parameters and data outside the normal ranges in order to uncover errors that may be hidden under standard parametrization. For example, with parameter “social dynamism” at 0, some entrepreneur-opportunities “rockets” entities may appear in the world. It is not a bug: it is due that some of the opportunities may coincide in the same spatial patch with an entrepreneur, and, therefore, if their state variables match, they become an entrepreneur-opportunities “rockets” entity “on the spot”.

However, the baseline code has not yet been peer-reviewed by other Netlogo modelers (except the supervisory team).

Graphical interface:



The interface has a set-up area, with slides, in which researchers can easily introduce the number of entrepreneurs, opportunities, financial institutions, workforce agencies, and clients for the parametrization and calibration of their empirical dataset under study. The ranges of the current baseline model can be change just clicking on the slide, and “Edit”.

The second set-up area consists in two slides with global parameters that affect all mobile agents, “search-cost” and “social-dynamism”.

The third set-up area has three buttons: “Set-up”, “Go once”, “Go”.

There are two plots with the histograms of the value of variables cash-flow and number of employees. Their code can be accessed right clicking and “Edit”. These two histograms show the heavy tailed distribution patterns in the majority of the runs (for the statistical analysis, see section “Model analysis” below).

There is also a plot called “populations” with the number of mobile agents and total number of employees. The exact numbers are counted in the monitors on the left of this plot.

In the interface, there is an “R-analysis” button that generates two CVS files with the distributions of cash-flows and number of employees at the end of the run, that can be easily imported into the statistical software R (or other mathematical software such as Matlab, Mathematica, etc.). We offered detailed software scripts and procedures below to analyze the distributions generated by the model.

MODEL OUTPUT VERIFICATION

The development of a model tries to reproduce some patterns of the real system, in this case, the emergence of heavy tailed distributions in nascent entrepreneurial outcomes, such as in the nascent firms' cash-flows or number of employees. In this TRACE section, we will offer detailed procedures and software packages in R to **quantitatively** analyse and decide if the obtained model outcome is a good enough representation of the heavy tailed distribution founded in the real longitudinal panels. The more observed patterns a model can reproduce at the same time, the more probability that it has captured the mechanisms of the real process satisfactorily well ('pattern-oriented modelling'; Grimm et al., 2005; Grimm and Railsback, 2012).

This section, output verification, is associated to what we previously defined as "face validation". Face validity illustrates that the processes and outcomes of the model are reasonable and plausible within its theoretical framework and the current knowledge in the research community.

Here a reminder of face validation concepts already described above:

- *Face validation*: the mechanism and properties of the model look like mechanisms and properties of reality. *Prima facie* (without detailed analysis) the model can convince that it contains elements and components that correspond to agents and mechanisms of the real world.
- *Empirical validation*: the model generates data that correspond to similar patterns of data in the real world. Data produced by the model must correspond to empirical data of the studied system. Empirical validation, therefore, often implies statistical tests and comparison between data sets. One of the problematic aspects of this type of validation is that real data is frequently with "noise", difficult to obtain, and partial. On the other hand, reality is not a computational machine with precise and well-defined results, but rather it yields messy results and it is very challenging to isolate and

measure the parameters of the real world. In this context, **calibration** is the process of finding the parameters and initial conditions that makes the model to match up as close as possible to the real, empirical datasets.

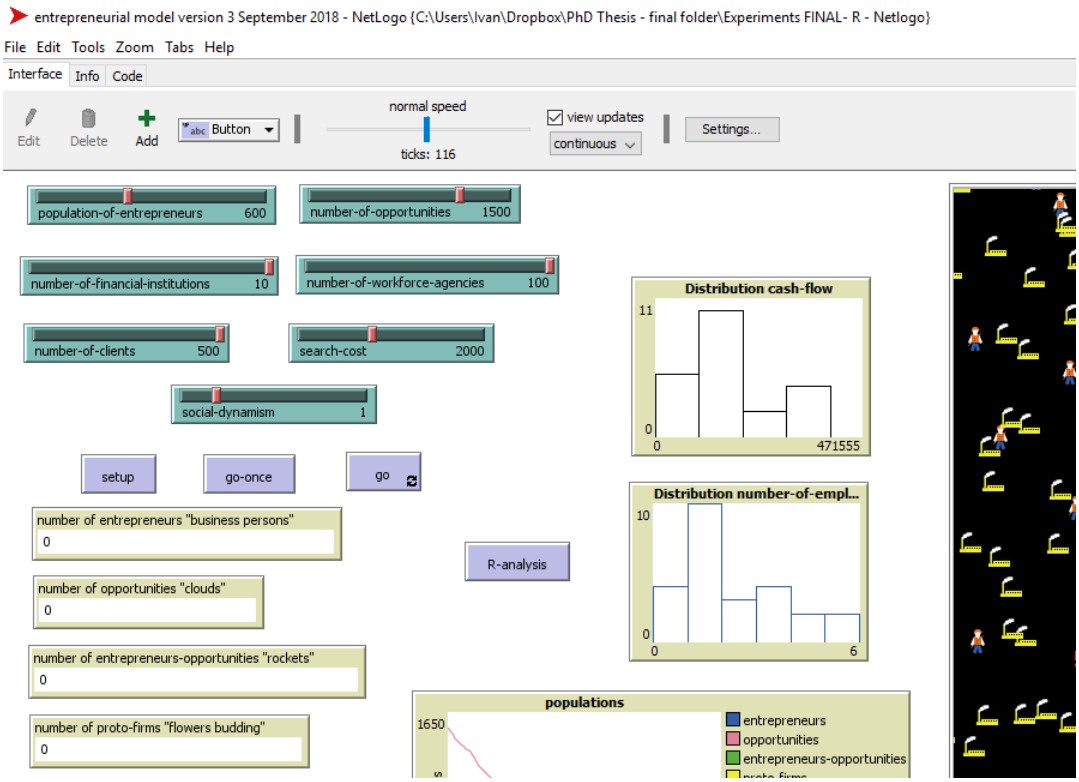
Regarding *face validation*, the "nascent entrepreneurial agent-based model" tries to simulate the entrepreneur's steps according to the conceptual framework currently accepted in the field (specifically Gartner, 1985, and Yang and Chandra, 2013).

Regarding the *empirical validation*, we will propose several formal tests that are based on multiple quantitative standards for a model matching a dataset (Railsback and Grimm, 2012, Chapter 20.4.2). However, we should notice that this model is purely stochastic, and that its results are conditioned by a multiplicative process: a small change in a certain value, in just one step, may vary the final results completely (path dependency). Therefore, we cannot expect similar patterns to those to the empirical datasets *in each and every run*. Because of the stochastic nature of the model and its multiplicative design some runs may not even generate results at all, especially when using extreme parameter values or swapping the conditions radically. The point of this section is to demonstrate that the model is able to generate distributions similar to reality - using parameters inspired on the real systems - and how to verify statistically this fact.

Next, we will describe an example of the procedure to analyse one random distribution generated by our model. As it has been said before, The "Nascent Entrepreneurial Agent-based model" is purely stochastically designed. Not every run, under every conditions or parameter, can match exactly with a determined and well-defined heavy tailed distribution. Many times, the obtained dataset can be plausibly fitted with several candidates (or log-normal or Pareto or Weibull, or all three). Real, empirical datasets are "messy" – difficult to identify -. The same happens with the results of the simulations.

Many runs of our model generate distributions that can be fitted as lognormal or power law with an exponential cut-off without any distribution identification problem. However, many other model distributions are in a statistical "twilight zone", in which two or more heavy tailed distributions can be considered a good fit. For this section, we have chosen two examples of "twilight" distributions, initially very difficult to identify, to show the reader the procedure for pitting using R statistical and fitting packages.

The first step is to check for an initial visual recognition (face validation) of heavy tailed distributions in the histograms located in the Netlogo interface, similar to those founded in empirical nascent entrepreneurial datasets. We may need to run the model several times because the model is stochastic. To analyses several runs, we will use the Netlogo tool "BehaviorSpace" (see below).



Once we have identified a good candidate run, we send the data to analysis pressing the button on the interface “R-analysis”. The button initiates a procedure that generates two CVS files with the last results of the run for the variables “cash-flow” and “number of employees” of the nascent firms.

The dataset files (in CVS) are located in the same folder where Netlogo is saved. Therefore, we need to establish the absolute path to the folder where the NetLogo in installed, starting from the root. On Windows, for example, something like “C:/Users/Ivan/Dropbox/PhD Thesis - final folder/Experiments FINAL- R - Netlogo/distributioncashflow.cvs”.

Here an example of R script for the testing procedure of the cash-flow distribution:

Import the CVS file into R:

```
R>library(readr)

R>distributioncashflow <- read_csv("C:/Users/Ivan/Dropbox/PhD Thesis - final folder/Experiments FINAL- R - Netlogo/distributioncashflow.cvs", col_names = FALSE, na = "empty")

R>View(distributioncashflow)
```

Initial tests using package ‘*goft*’ version 1.3.4

```
R>install.packages(goft)

R>library(goft)

R>lnorm_test(distributioncashflow$X1[distributioncashflow$X1>=0])
```

(Notice that we have to take only positive values because zeros or negative values may produce mathematical indetermination for some script calculations. Also notice that some tests for these positive heavy tailed distributions cannot even deal with 0 values in the variable set).

Test for the lognormal distribution based on a transformation to normality

data: distributioncashflow$X1[distributioncashflow$X1 >= 0]
p-value = 0.03536

R>gp_test(distributioncashflow$X1[distributioncashflow$X1 >= 0])

Bootstrap test of fit for the generalized Pareto distribution

data: distributioncashflow$X1[distributioncashflow$X1 >= 0]
p-value = 0.5275

R>weibull_test(distributioncashflow$X1[distributioncashflow$X1 >= 0])

Test for the Weibull distribution

data: distributioncashflow$X1[distributioncashflow$X1 >= 0]
p-value = 0.98

R>gamma_test(distributioncashflow$X1[distributioncashflow$X1 >= 0])

Test of fit for the Gamma distribution

data: distributioncashflow$X1[distributioncashflow$X1 >= 0]

V = -1.1805, p-value = 0.4038

normal_test(distributioncashflow$X1[distributioncashflow$X1 >= 0])

Correlation test for normality

data: distributioncashflow$X1[distributioncashflow$X1 >= 0]
R = 0.99369, p-value = 0.1051
Alternative hypothesis:
distributioncashflow$X1[distributioncashflow$X1 >= 0] does not follow a normal distribution.

Tests run by package 'goft' point out to the Weibull as the best fit (p-value = 0.98)

Tests with package 'fitdistrplus' version 1.0-9

R>install.packages(fitdistrplus)
R> library("fitdistrplus")

To fit a distribution to a dataset is normally needed to choose good distribution candidates among the plausible ones. We choose these candidates based of the knowledge of the processes governing the variable to be modeled – multiplicative processes in our model - or by the observation of its plot – like we did looking at the histograms of the interface -.

Package "**fitdistrplus**" offers another tool to help this initial choice, the function "**plotdist**" that plots the distribution and its density, and the cumulative distribution function (CDF).

R>plotdist(distributioncashflow$X1[distributioncashflow$X1 >= 0], histo = TRUE, demp = TRUE)

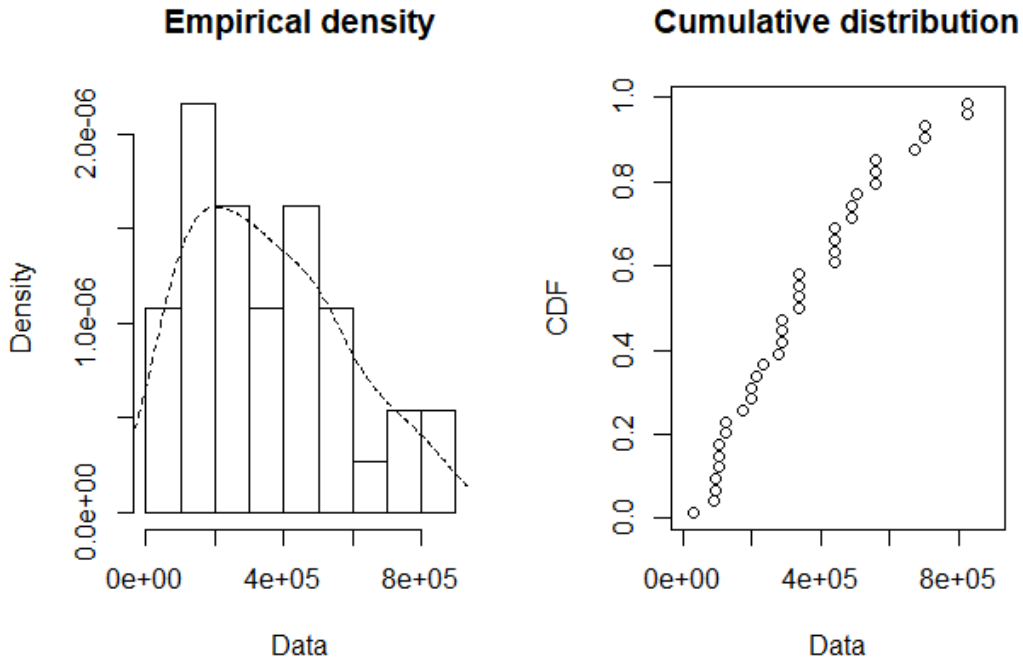
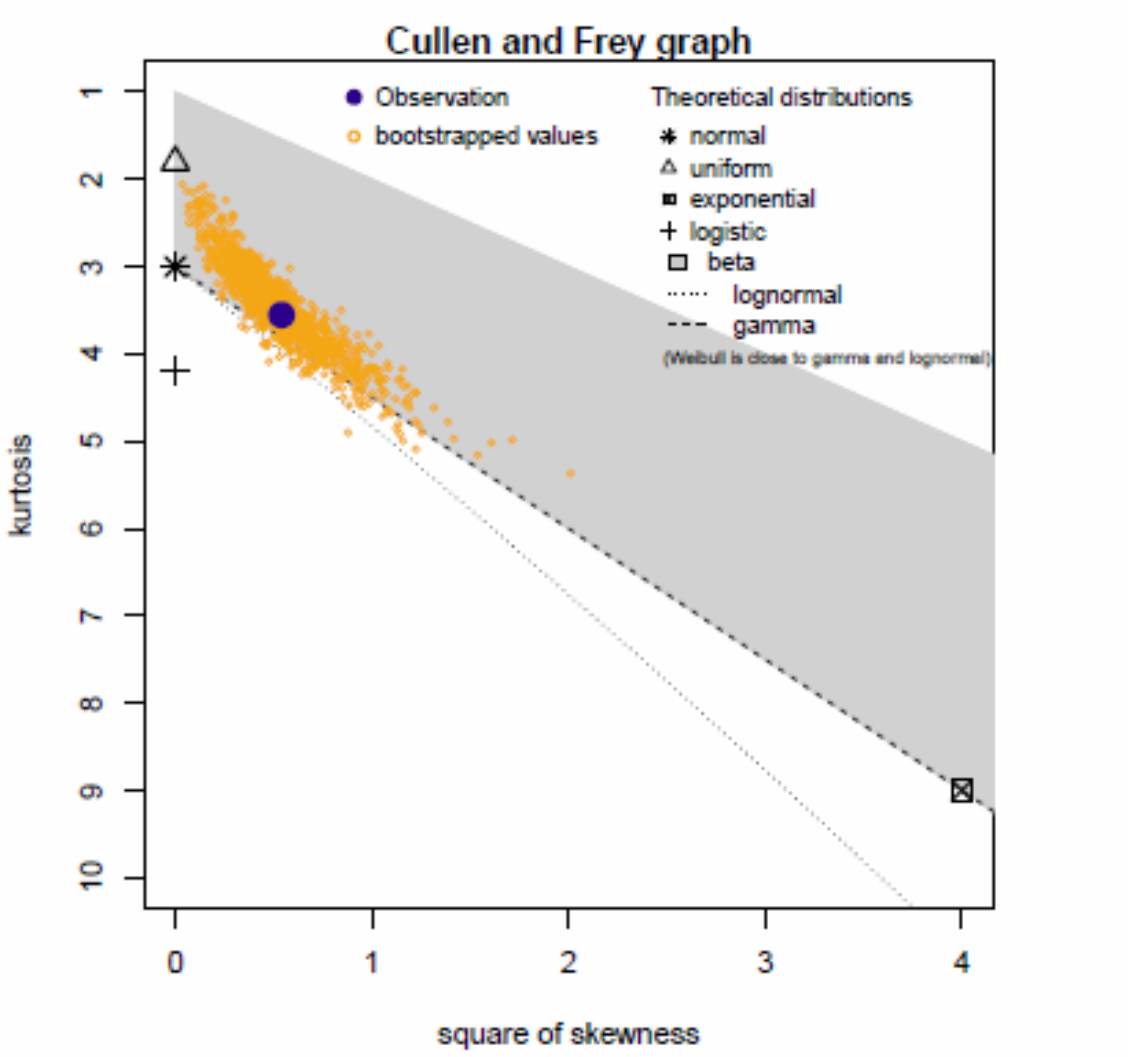


FIGURE 23 - HISTOGRAM AND CDF PLOTS OF THE CASH-FLOW DISTRIBUTION AS PROVIDED BY THE PLOTDIST FUNCTION.

Another useful function in "**fitdistrplus**" is **descdist**, which provide an indicative skewness-kurtosis plot that can help to identify the best candidates.

R>descdist(distributioncashflow$X1[distributioncashflow$X1 >= 0], boot = 1000)


Cullen and Frey graph
Observation
bootstrapped values
Theoretical distributions
normal
uniform
exponential
logistic
beta
lognormal
gamma
(Weibull is close to gamma and lognormal)
kurtosis
square of skewness
0
1
2
3
4
1
2
3
4
5
6
7
8
9
10

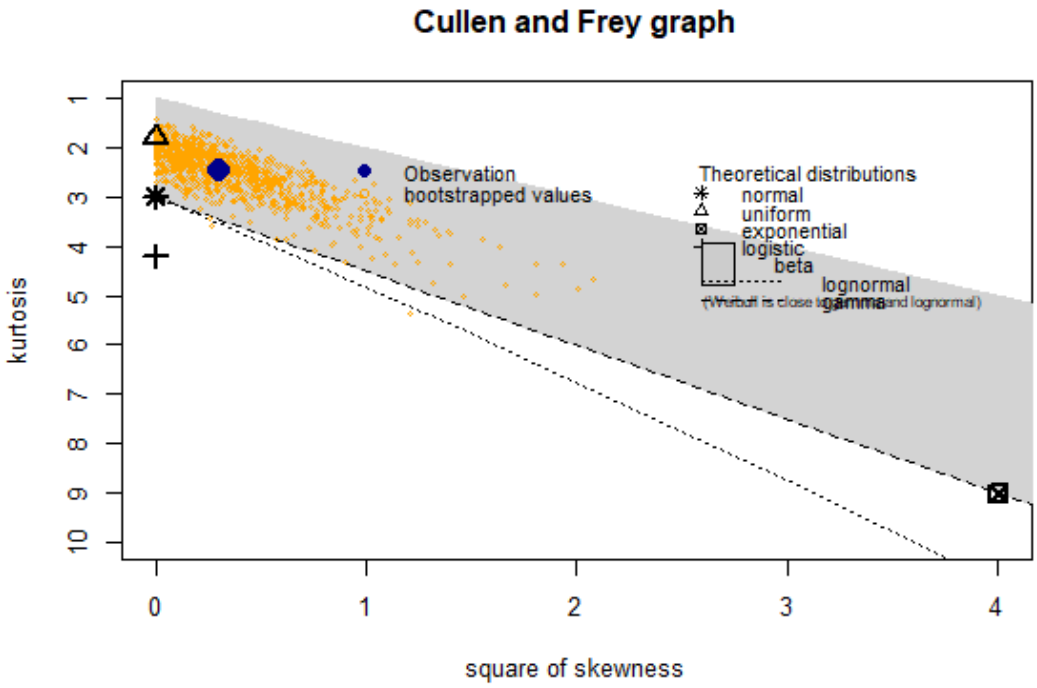


FIGURE 24 - EXAMPLES OF SKEWNESS-KURTOSIS PLOTS FOR AS PROVIDED BY THE DESCDIST FUNCTION. THE FIRST FIGURE IS FROM THE R PACKAGE (DELIGNETTE-MULLER AND DUTANG, 2015). THE SECOND FIGURE IS ONE EXAMPLE OF CASH-FLOW DISTRIBUTION GENERATED BY OUR MODEL.

Looking at the plots of these examples with a positive skewness and a kurtosis close to 3, the fit of the three more common right-skewed distributions can be considered, that is, Weibull, gamma and lognormal distributions.

The next step is to generate the goodness-of-fit plots. This procedure is performed by the function **fitdist** and it offers four goodness-of-fit plots (Cullen and Frey, 1999; Delignette-Muller and Dutang, 2015):

- A density plot with the density function of distribution and its histogram.

- A CDF plot of both the distribution under study and the fit of the candidate distribution (Weibull, log-normal, Pareto, etc.).

- A Q-Q plot representing the distribution quantiles (y-axis) against the fitted quantiles (x-axis).

- A P-P plot with distribution function evaluated at each data point (y-axis) against the candidate distribution function (x-axis).

Fitting Weibull

R> fw <- fitdist(distributioncashflow$X1[distributioncashflow$X1 >= 0], "weibull")

R>plot(fw)

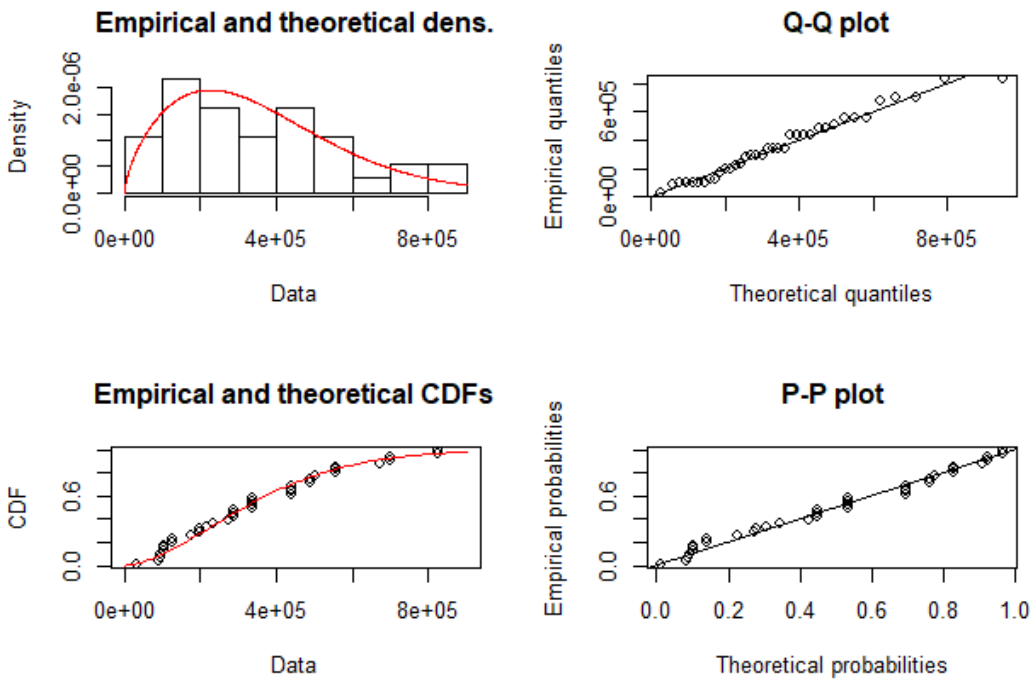


FIGURE 25 – FITTING WEIBULL OF OUR SAMPLE

R>summary(fw)

Fitting of the distribution ' weibull ' by maximum likelihood

Parameters :

	estimate	Std. Error
shape	1.658111e+00	0.2077179
scale	3.937846e+05	8407.7671280

Loglikelihood: -503.6075 AIC: 1011.215 BIC: 1014.437

Correlation matrix:

	shape	scale
shape	1.00000000	0.06747066
scale	0.06747066	1.00000000

Fitting a gamma distribution

R>fg <- fitdist(distributioncashflow$X1[distributioncashflow$X1 >= 0], "gamma")

Although this distribution is considered a potential candidate, the characteristics of our distribution dataset do not allow the algorithms to generate the gamma fitting. The R script produces an error message.

Fitting a lognormal distribution.

R>fln <- fitdist(distributioncashflow$X1[distributioncashflow$X1 >= 0], "lnorm")

R>plot(fln)

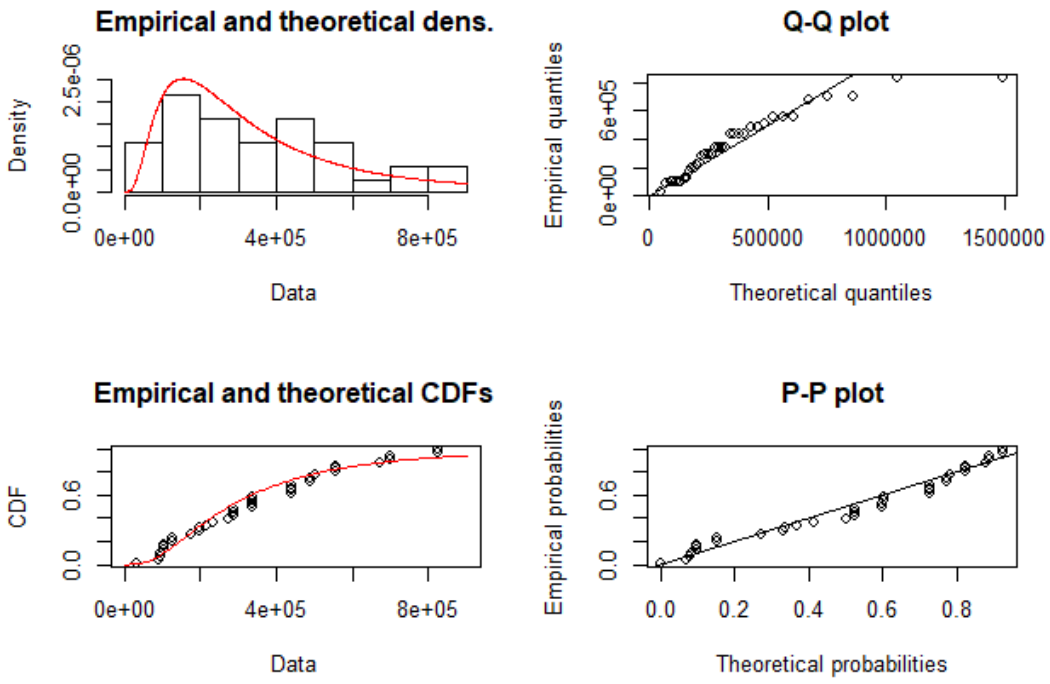


FIGURE 26 – FITTING A LOGNORMAL DISTRIBUTION

Fiting a beta distribution.

R>flb <- fitdist(distributioncashflow$X1[distributioncashflow$X1 >= 0], "beta")

Although this distribution is considered a potential candidate, the characteristics of our distribution dataset do not allow the algorithms to generate the beta fitting. The R script produces an error message.

Comparison between lognormal and Weibull candidate distributions and their plots (defined above):

R>par(mfrow = c(2, 2))

R>plot.legend <- c("Weibull", "lognormal")

R>denscomp(list(fw, fln), legendtext = plot.legend)

R>qqcomp(list(fw, fln), legendtext = plot.legend)

R>cdfcomp(list(fw, fln), legendtext = plot.legend)

R>ppcomp(list(fw, fln), legendtext = plot.legend)

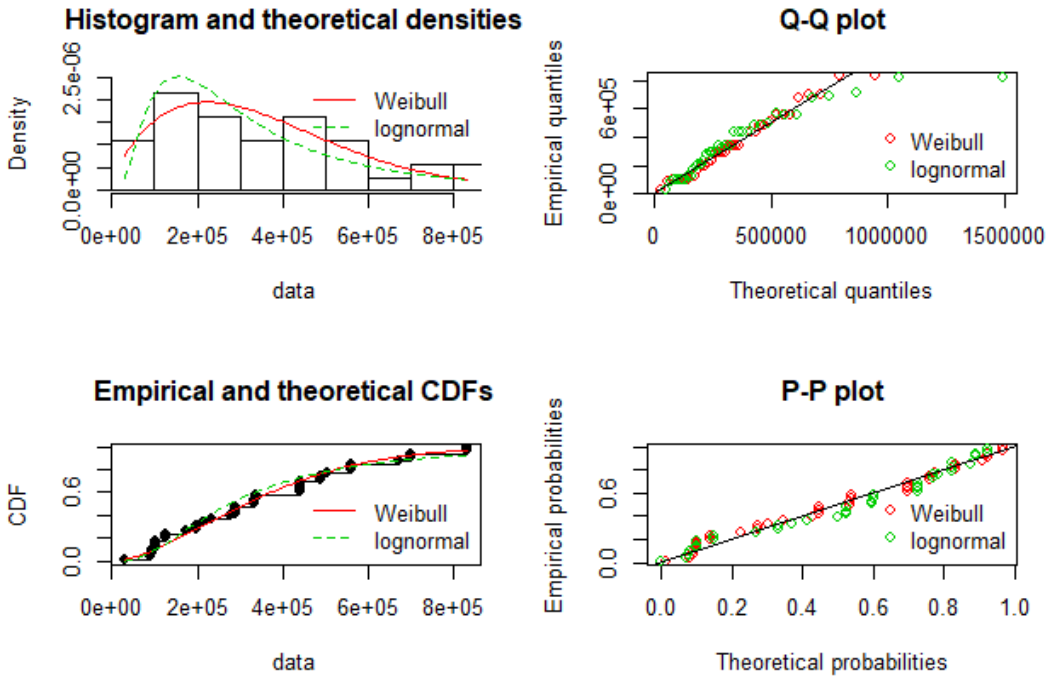


FIGURE 27 - COMPARISON BETWEEN LOGNORMAL AND WEIBULL CANDIDATE DISTRIBUTIONS AND THEIR PLOTS

Goodness-of-fit

The **fitdistrplus** R package computes different goodness-of-fit statistics in order to compare among the candidate fitted distributions. These goodness-of-fit statistics measure the distance between the proposed fitted distribution and our model distribution, that is, the distance between the fitted cumulative distributions of the candidate distribution with the cumulative distribution of our dataset. **Fitdistrplus** considers three classic statistics (D'Agostino and Stephens, 1986; Delignette-Muller & Dutang, 2015):

- Cramer-von Mises statistics
- Kolmogorov-Smirnov statistics
- Anderson-Darling statistics

This comparison procedure is performed by the function **gofstat**, as developed by Stephens (D'Agostino and Stephens, 1986; Delignette-Muller & Dutang, 2015):

```
R>gofstat(list(fw, fln), fitnames = c("weibull", "lnorm"))
```

Goodness-of-fit statistics

	weibull	lnorm
Kolmogorov-Smirnov statistic	**0.10418304**	0.1328885
Cramer-von Mises statistic	**0.05339458**	0.1181312
Anderson-Darling statistic	**0.38414190**	0.7467988

Goodness-of-fit criteria

	weibull	lnorm
Akaike's Information Criterion (AIC)	**1011.215**	1016.197
Bayesian Information Criterion (BIC)	**1014.437**	1019.418

Although we cannot reject the candidacy of the lognormal distribution, the smaller distance in all goodness-of-fit correspond to the Weibull distribution, and, therefore, we should consider the Weibull distribution a better fit.

Similarly, we can follow the same procedure to analyze the distribution of the variable “number of employees” at the end of the run. Again, given the stochastic nature of the model, within the same run, the distribution of one variable may not coincide with the best fit of other variables of the model. For example, the cash-flow distribution can be clearly identified as a type of a heavy tailed distribution, and, however, even in the same run, another variable, for example, “number of employees” may fit another distribution better or it may not follow any defined pattern at all. We show an example of this type of behavior in the following procedure. Data were collected for the same run. We analyzed cash-flow above (Weibull and log-normal were good fit). Below we follow the same procedure for

the variable “number of employees” of the same run. Again, we have chosen a “problematic” simulation distribution to illustrate how we can find the best fit.

```
R>library(goft)
R>library(fitdistrplus)

R>library(readr)
R>distributionofemployees <- read_csv("absolute path to cvs file",
   col_names = FALSE, na = "empty")
R>View(distributionofemployees)
```

Goft package analysis

```
R>lnorm_test(distributionofemployees$X1[distributionofemployees>0])
```

(Notice that we have to remove the zeros because they produce mathematical indetermination in the script calculation. Many tests for these positive heavy tailed distributions cannot deal with 0 values in the variable set).

Test for the lognormal distribution based on a transformation to normality

```
data:  distributionofemployees$X1[distributionofemployees > 0]
p-value = 0.0008227
```

```
R>gp_test(distributionofemployees$X1[distributionofemployees>0])
```

Bootstrap test of fit for the generalized Pareto distribution

```
data:  distributionofemployees$X1[distributionofemployees > 0]
p-value = 0.3023
```

R>weibull_test(distributionofemployees$X1[distributionofemployees>0])

Test for the Weibull distribution

data: distributionofemployees$X1[distributionofemployees > 0]
p-value = 0.678

R>gamma_test(distributionofemployees$X1[distributionofemployees>0])

Test of fit for the Gamma distribution

data: distributionofemployees$X1[distributionofemployees > 0]
V = -1.3892, p-value = 0.326

R>normal_test(distributionofemployees$X1[distributionofemployees>0])

Correlation test for normality

data: distributionofemployees$X1[distributionofemployees > 0]
R = 0.99548, p-value = 0.2141

Alternative hypothesis:
distributionofemployees$X1[distributionofemployees > 0] does not follow a normal distribution.

Tests with the R package "fitdistriplus"

R>plotdist(distributionofemployees$X1[distributionofemployees>0], histo = TRUE, demp = TRUE)

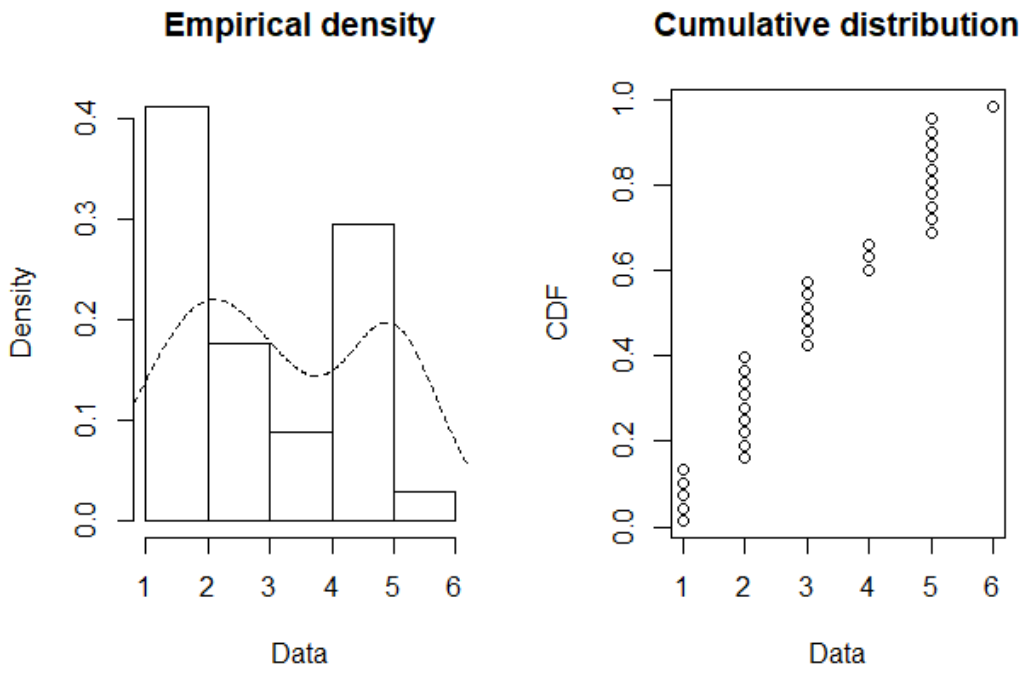


FIGURE 28 – PLOTDISC OF "NUMBER OF EMPLOYEES" SIMULATION: HISTOGRAM AND DENSITY FUNCTION

From the **plotdist** histogram and density function, we can clearly foresee the challenge of finding a good fit.

```
R>descdist(distributionofemployees$X1[distributionofemployees>0], boot = 1000)
    summary statistics
    ------
    min:  1   max:  6
    median:  3
    mean:  3.205882
    estimated sd:  1.552699
    estimated skewness:  0.0994631
    estimated kurtosis:  1.602816
```

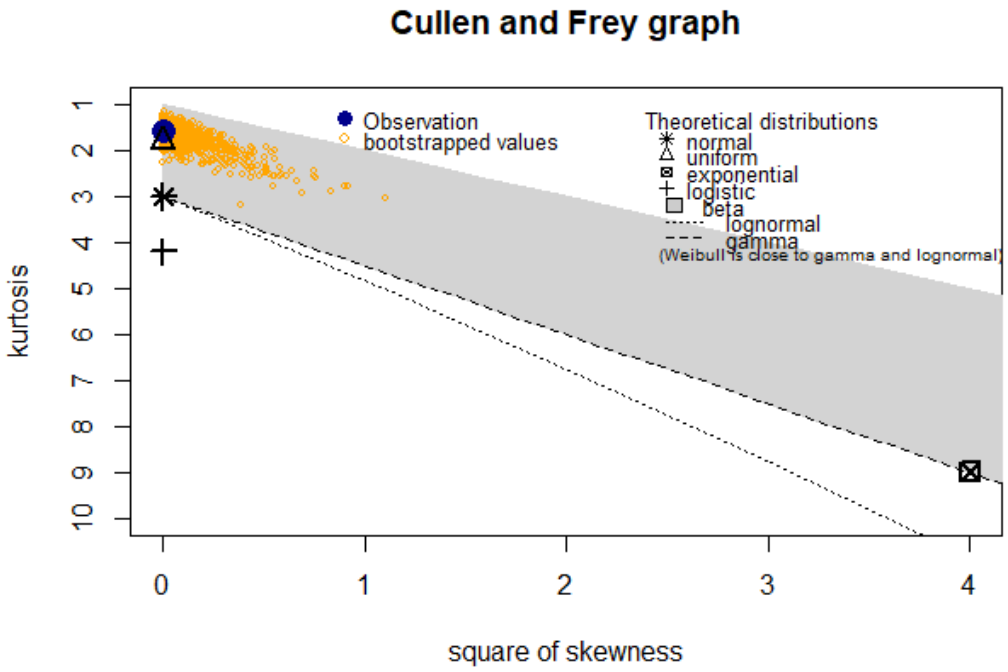


FIGURE 29 – CULLEN AND FREY GRAPH OF "NUMBER OF EMPLOYEES" SAMPLE

Fitting a Weibull distribution:

R>fw <- fitdist(distributionofemployees$X1[distributionofemployees>0]], "weibull")

R>plot(fw)

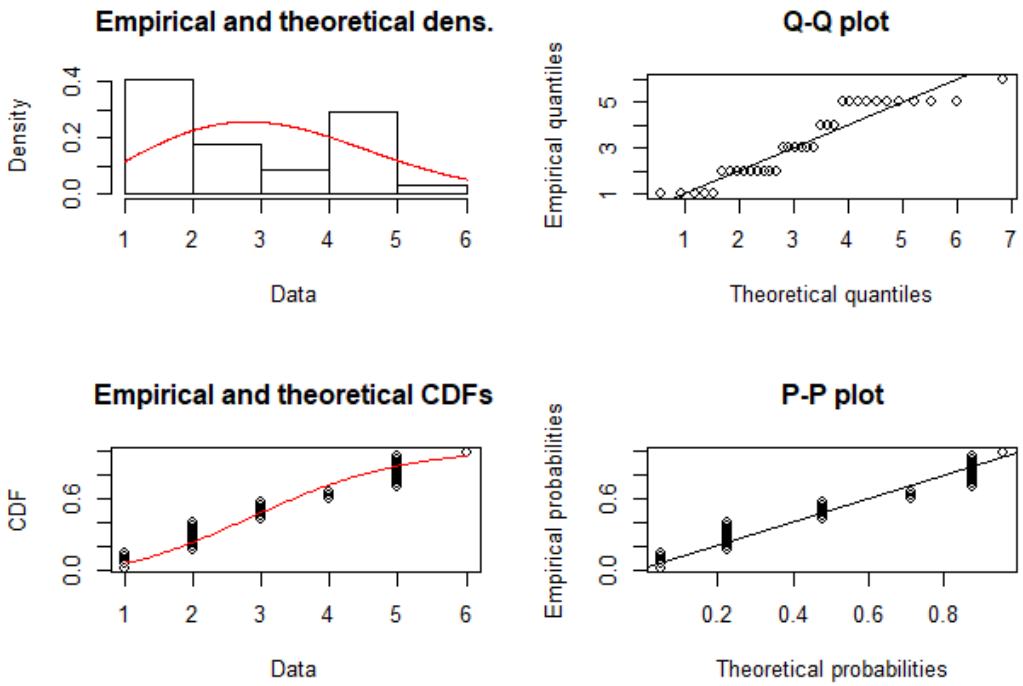


FIGURE 30 - FITTING A WEIBULL DISTRIBUTION

R>summary(fw)

Fitting of the distribution ' weibull ' by maximum likelihood

Parameters :

	estimate	Std. Error
shape	2.264646	0.3176993
scale	3.630203	0.2894863

Loglikelihood: -61.37971 AIC: 126.7594 BIC: 129.8121

Correlation matrix:

	shape	scale
shape	1.0000000	0.3135553
scale	0.3135553	1.0000000

Fitting a gamma distribution:

R>fg <- fitdist(distributionofemployees$X1[distributionofemployees>0], "gamma")

R>plot(fg)

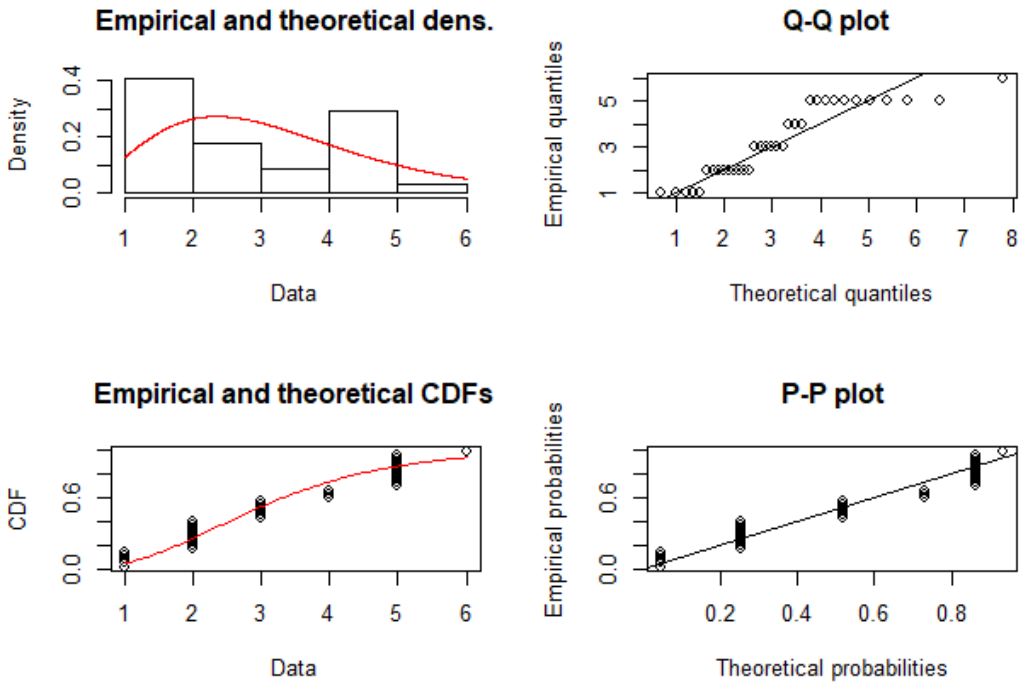


FIGURE 31 - FITTING A GAMMA DISTRIBUTION

R>Summary(fg)

Fitting of the distribution ' gamma ' by maximum likelihood

Parameters :

	estimate	Std. Error
shape	3.749289	0.8719036
rate	1.169623	0.2910570

Loglikelihood: -62.15744 AIC: 128.3149 BIC: 131.3676

Correlation matrix:

	shape	rate
shape	1.0000000	0.9345161
rate	0.9345161	1.0000000

Fitting a lognormal distribution:

R>fln <- fitdist(distributionofemployees$X1[distributionofemployees>0], "lnorm")

R>plot (fln)

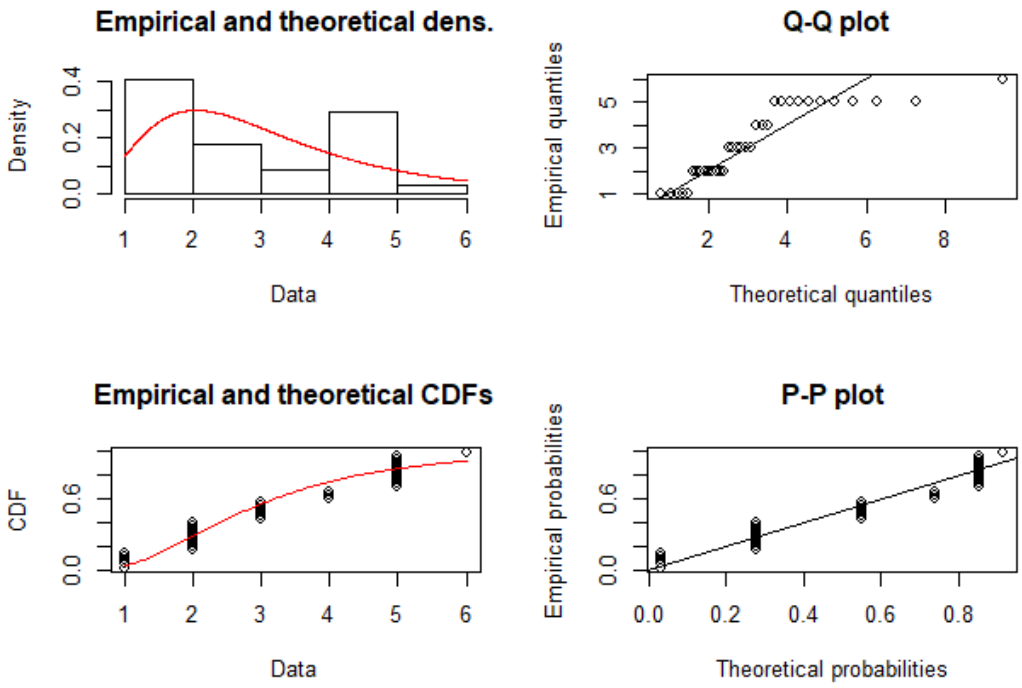


FIGURE 32 - FITTING A LOGNORMAL DISTRIBUTION

R>summary(fln)

Fitting of the distribution ' lnorm ' by maximum likelihood

Parameters :

	estimate	Std. Error
meanlog	1.0257359	0.09608222
sdlog	0.5602508	0.06793942

Loglikelihood: -63.42033 AIC: 130.8407 BIC: 133.8934

Correlation matrix:

	meanlog	sdlog
meanlog	1	0
sdlog	0	1

Beta Fitting

flb <- fitdist(distributionofemployees$X1[distributionofemployees>0], "beta")

(Although this distribution is considered a potential candidate, the characteristics of our distribution dataset do not allow the algorithms to generate the beta fitting. The R script produces an error message).

Comparison between Weibull, gamma and lognormal distributions:

R>par(mfrow = c(2, 2))

R>plot.legend <- c("Weibull", “gamma”, "lognormal")

R>denscomp(list(fw, fg, fln), legendtext = plot.legend)

R>qqcomp(list(fw, fg, fln), legendtext = plot.legend)

R>cdfcomp(list(fw, fg, fln), legendtext = plot.legend)

R>ppcomp(list(fw, fg, fln), legendtext = plot.legend)

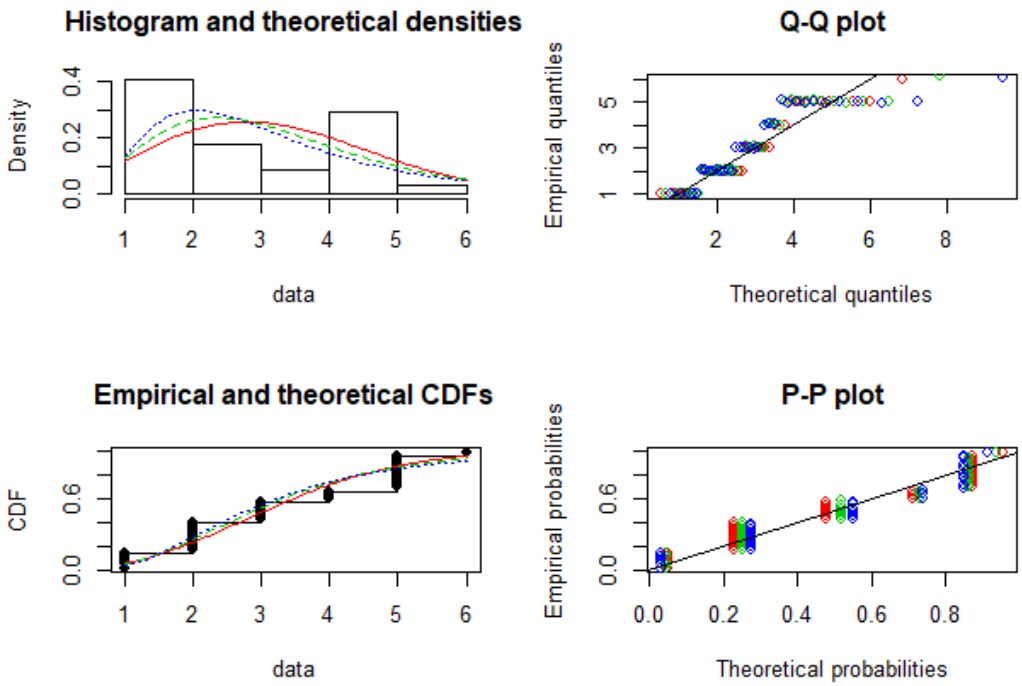


FIGURE 33 - COMPARISON BETWEEN WEIBULL, GAMMA AND LOGNORMAL DISTRIBUTIONS

R>gofstat(list(fw,fg, fln), fitnames = c("weibull", "gamma", "lnorm"))

Goodness-of-fit statistics

	weibull	gamma	lnorm
Kolmogorov-Smirnov statistic	0.1966836	0.186984	0.**1747907**
Cramer-von Mises statistic	0.2245667	0.**2118663**	0.2229274
Anderson-Darling statistic	1.5239191	1.**4793666**	1.5990575

Goodness-of-fit criteria

	weibull	gamma	lnorm
Akaike's Information Criterion	**126.7594**	128.3149	130.8407
Bayesian Information Criterion	**129.8121**	131.3676	133.8934

The best fit tests are not conclusive given the divergence of the different goodness-of-fit statistics and criteria; however, if we take into consideration the

goft test, the Weibull distribution would be a good candidate with a higher p value (p-value = 0.678).

This example shows the difficulties and challenges to find the best fit distribution for a dataset, either from a real system or from a simulation with a high level of stochasticity.

To generate multiple runs, Netlogo provides a useful tool that allows to export the outcomes called BehaviorSpace. The script is already coded and implemented in the model (go to the "Tools" label → BehaviorSpace → experiment "Behavior Space Output Verification").

As before, we have considered parameters resembling Swedish conditions. However, this should not be considered a "calibration" of the model. As a research tool, the "nascent entrepreneurial agent-based model" has been designed to allow more detailed parametrization and calibration with the different longitudinal panel worldwide. It is, therefore, generic and flexible at this point of development.

["number-of-opportunities" 1500]
["social-dynamism" 1]
["number-of-workforce-agencies" 100]
["population-of-entrepreneurs" 600]
["number-of-clients" 500]
["search-cost" 2000]
["number-of-financial-institutions" 10]

We have coded a "Final Command" script that put together in a CVS file the results of the multiple runs:

Netlogo code:

```
;;File with cash-flows of new firms;;;;;;
```

```
file-open "behaviourspacecashflow.cvs"
 ask new-firms
 [
  file-print (last [cash-flow] of new-firms)
  ]

file-print "end of the run" ;; to know when the run ends

 file-close



 ;;;Now [number-of-employees] of new-firms;;;;;;;;;;;;;;;;

file-open "behaviourspaceemployees.cvs"
 ask new-firms
 [
    file-print ( last [number-of-employees] of new-firms)
  ]

file-print "end of the run" ;; to know when the run ends

file-close
```

MODEL ANALYSIS

This section of the TRACE document is related to the question: Can we still verify the results of the model if we introduce small changes in one or two parameters? We can study the model deeper by performing controlled simulation experiments keeping some parameters constant and changing one or more over a wider range. Then, we can explore the consequences of these variations in the distribution of the output variables. Local sensitivity analysis help us to understand and evaluate how sensitive are the outputs - the variables distributions - to small changes in one parameter at a time.

The model analysis should also include experiments with simplified versions of the model, in which the "world" in which the agents behave is more homogenous and constant, with reduce system size, and in which certain processes are deactivated. **Our model is currently in this stage.** Although complex features have been coded, presently, many of them have been muted - with the muting sign in Netlogo " ; " - for simplification purposes such as the "opportunities-generators", the entrepreneurs' teams formation, the entrepreneur-financial-resources assignment (currently randomly assigned instead of through a defined distribution), more complex environmental variables (time series), etc.

Sensitivity analysis should also be performed on initial conditions and input data. For example, when the model is calibrated, the input data will correspond to the information provided by the panels on each of the entrepreneurs (capital resources, etc.). The input data code has also been implemented in the "Nascent Entrepreneurial agent-based model", in the procedure "to setup-entrepreneurs", although now it is muted (input file "entrepreneur-financial-resources-EmpiricalData.txt").

Although currently the model is in the simplified version, the parameter space is so huge that, at this point, is not feasible to offer - under the scope of this PhD research - a comprehensive sensitivity analysis due to the relatively high computation times. For example, the study of the sensitivity analysis of just the parameters located in the interface would generate this number of possible states:

Parame-ter	Population Of Entrepre-neurs	Number of Opportu-nities	Number of Financial institutions	Number of Workforce Agencies	Number of clients	Search cost	Social Dyna mism	Total
Range	0-1500	0-2000	0-10	0-100	500	0-5000	0-5	
Parame-ter Space	1500	2000	10	100	500	5000	5	$3.75*10^{16}$

Obviously, we do not need to explore the complete set of $3.75*10^{16}$ possibilities: only few areas of this parameter space are able to generate heavy tailed distributions. In complex models like the "Nascent Entrepreneurial Agent-based model", computational run time, complexity and stochasticity will limit global sensitivity analysis, and only a subset of parameters could be realistically analyzed.

The sensitivity analysis should focus on those parameters more uncertain (such as the number of opportunities), or some of the parameters included in the code that were heuristic or that are very difficult to know their value in the real system (financial institution criteria for investment, characteristics of the business opportunities, etc.). Thus, the sensitivity analysis would also offer conclusions regarding the model uncertainty. If the model is very sensitive to the parameters that are more uncertain, then, the entire model should be considered quite uncertain. At the contrary, if the model is less sensitive to the most uncertain parameters, the model will have more possibility to pass the uncertainty analysis

tests. In any case, only the researchers working with the empirical results of the PSED-like longitudinal panel are in the position to know and to decide which the more uncertain parameters are. These uncertain parameters may be different in each case and for very diverse reasons (wrong design of the panel or questions, interviewers' mistakes, database errors or "crashes", missing values, "non-disclosure" of the participants, etc.).

On the other hand, the sensitivity analysis would only make sense when the model is fully parametrized and calibrated for a specific, concrete longitudinal panel dataset. Each country has its own specificities and the impact of the variation of one parameter may be different depending on the country under study. The sensitivity analysis will indicate which processes are the most important for obtaining the heavy tailed distributions observed in the empirical datasets.

We should notice that parameters are not mere numbers obtained from the empirical datasets of the longitudinal panels. Often, they represent entire processes that we, as modelers, decided not to represent explicitly. For example, our entrepreneur's variable "capacity-to-achieve", expressed by a percentage, is a numeric representation of the entrepreneur´s social and human capital, strong and weak tie networks, acquaintance with investment capital, opportunity recognition capabilities, entrepreneur's education, previous experience in industry or venture founded, genetic factors, etc. The complexity of this variable is so huge that we have decided to agglomerate all the factors in just a percentage.

Similarly, submodels also represent processes that are represented explicitly in more details, but that still are a coarse simplification of reality. Therefore, submodels should also be analyzed by contrasting alternative submodels, or even, changing the order of the procedures. For example, in our model, how would the results change if the entrepreneur first search for employees (building the team first) instead of searching first for financial institutions (money)?

Making the submodels more complex or simpler may provide relevant insights into our model design. For example, a sensitivity analysis of our submodels should study the impact of generating some of the state variable of the agents following a specific distribution function (Poisson, lognormal, etc.) instead of the current exclusive random generation.

Taking again a practical approach, we propose here to the reader the better tools regarding how to implement a sensitivity analysis for the "nascent entrepreneurial agent-based model" once it has been adapted the parameters to a specific empirical longitudinal panel dataset. Those procedures also would help to calibrate properly the model.

The main reference to address a sensitivity analysis with a Netlogo model using R is the work of Thiele (2010, 2012, 2014):

> Thiele, J.C., Kurth, W. and Grimm, V., 2014, Facilitating Parameter Estimation and Sensitivity Analysis of Agent-Based Models: A Cookbook Using NetLogo and 'R'. *Journal of Artificial Societies and Social Simulation,* vol. 17, no. 3.
>
> http://jasss.soc.surrey.ac.uk/17/3/11.html

They offer a quite comprehensive "cookbook" to perform the calibration and sensitivity analysis. The complete set of scripts is in the "Supplementary Materials" located in this repository:

> http://sourceforge.net/projects/calibrationsensitivityanalysis/

The scripts are quite straightforward for each method: we only have to change the variables names and some values and tests, and to understand the procedure. A warning: the analytical procedures require a solid background in R programming and advance statistics: sequential Monte Carlo (SMC) method, Evolutionary algorithms (EA), Bayesian methods, etc.

It will also require to install the interface package between Netlogo and R, mentioned in the previous section, **RNetlogo**. The best introductions and tutorials to learn this package are these:

Thiele, J.C., 2014. R Marries NetLogo: Introduction to the RNetLogo Package. *Journal of Statistical Software,* vol. 58, no. 1, pp. 1-41.

http://www.jstatsoft.org/v58/i02/

Thiele, J., Kurth, W., and Grimm, V., 2012. RNetLogo: An R package for running and exploring individual-based models implemented in NetLogo. *Methods in Ecology and Evolution*, 3(3), 480–483.

Thiele, J.C. and Grimm, V., 2010. NetLogo meets R: Linking agent-based models with a toolbox for their analysis. *Environmental Modelling and Software,* vol. 25, no. 8, pp. 972-974.

MODEL OUTPUT CORROBORATION

The "nascent entrepreneurial agent-based model" was designed for understanding the processes that occur in the emergence of nascent firms. As said above, the baseline model has not been parametrize and calibrate to any specific empirical longitudinal panel dataset yet. The parameters used in the baseline model code resemble the Swedish conditions, but many are still heuristic. The purpose was to show that the complex multiplicative processes related to the constitution of a nascent firm can be simulated by a complex agent-based model and that we can obtain heavy tailed distributions very similar to the empirical outcomes.

This section refer to the potential predictive possibilities of the model, that is, the capacity of the model to make predictions that can be confirm subsequently in the empirical setting. At this stage of the model development, we are not there yet. Once the model is parametrize and calibrate, we can identify the variables and parameters that are relevant for policy making. Indeed, two of the major objectives of entrepreneurial research are 1) to be able to build theory that help us to understand the birth of new firms and 2) to design policies and strategies for the foundations of solid entrepreneurial ecosystems.

Further developments, modifications and refinements of this model would help us to know which the key factors that may increase the number of new firms, their survival rates and the number of employees are. However, the model should be previously customized to each of the system under study because these factors can be different in each country or region. On the other hand, the complexity of the entrepreneurial process makes impossible that an agent-based model - or any other model technique - really captures the actual dynamics of the nascent entrepreneurial emergence sufficiently well. Prediction in complex system is still challenging and limited (weather forecast – meteorology -, earthquake forecast – geology -, financial crisis forecast – economics -, etc.). Agent-based modeling may

have good predictive capabilities in other fields, but its potential in entrepreneurship research still has to be elucidated. The purpose of our investigation is to open this line of research.

Model output verification consists in tuning the model parameters, environmental conditions and submodel designs to reproduce the empirical observations, that is, the outcome distributions. This is necessary because modelling requires compulsorily some kind of simplification of reality. Often we have to compensate for processes to complex to model, lack of sufficient information on the system under study, or the need of keeping the model simple enough to understand it and communicate it. **However, the real good indicator of structural realism of the model is only achieved when the model is able to predict phenomena that were not conceived during the development of the model and its testing.** This is what it is called "Model output corroboration". This standard is very difficult to obtain in many disciplines, for example, in climate change or ecology: experimentation is not always feasible or ethical. In entrepreneurship, it would require the involvement of not only the entrepreneurship research community but also of the rest of the stakeholders such as policy makers, institutions involved, entrepreneurs, etc.

8. CONCLUSIONS AND FUTURE RESEARCH

This research was done in the context of the on-going dialog and debate regarding the search for the generative processes in nascent entrepreneurship, and, more broadly, in the discovery of heavy-tail distributions in inputs and outcomes variables across different nascent entrepreneurial panel studies performed in different countries and continents (Andriani & McKelvey, 2009; Reynolds and Curtin, 2011; Crawford and McKelvey, 2012; Crawford et al., 2014; Crawford et al., 2015; Reynolds, 2017,b).

Joo, Aguinis and Bradley (2017) proposed a new distribution pitting methodology for the assessment of the types of non-normal distributions (Joo, Aguinis and Bradley, 2017). In the first section of this research, we followed their methodology for the analysis of the empirical nascent entrepreneurial outcomes in those countries in which the datasets are in the public domain: Australia, Sweden, US PSED I & II (Reynolds, 2017b). The implementation of the distribution pitting was through a new R statistical package, called **Dpit**.

After applying the **Dpit** statistical package to the outcomes variables of nascent entrepreneurial datasets, we found that the results mostly suggested two types of distributions for these entrepreneurial samples: power law with an exponential cut-off and lognormal distributions (occasionally, Weibull distributions would also be a good fit). However, the results were not completely conclusive. Which of these two distributions may be the better fit will require the analysis of the rest of 14 still ongoing longitudinal projects around the world. The pervasiveness of lognormality offers relevant clues to understand nascent entrepreneurial processes, their

generative mechanism, and it will offer strategies to allocate resources to foster and promote new entrepreneurial ventures.

The second objective of this research was the design and implementation (coding) of an agent-based model with enough complexity to be able to simulate the heavy tailed distributions patterns in the different international empirical longitudinal studies. It was conceived and intended as a research tool - openly available to the research community - to test and explore new theories and empirical datasets in nascent entrepreneurial processes.

Our “nascent entrepreneurial agent-based model”, inspired by previous simpler entrepreneurial models, introduces new layers of complexity, making possible parametrization and calibration (not possible in the previous seminal entrepreneurial agent-based model attempts). This baseline model, initially with parameters similar to the public available panel datasets --Australia, Sweden, US PSED **--, is able to generate the patterns** that were found in the empirical results: the heavy-tailed distributions.

This baseline model has a flexible design in order to be easily adapted to each of the empirical dataset under study. The model, at this initial stage, has not been fully parametrized and calibrated for any specific country. The baseline model takes the main parameters from the datasets available heuristically, in order to show that multiplicative processes --as main generative mechanism-- are able to simulate the empirical patterns.

The baseline model was designed as a research tool to experiment and to help entrepreneurship researchers to test their theories, and for exploring in more detail the mechanisms involved in the emergence of new ventures. The baseline model and its background documentation will be

openly available to the research community in two major agent-based repositories. Taking this baseline model as a “backbone”, researchers can change parameters, agents, behaviours, schedules or global variables for their own theory building or calibration of their specific country’s simulation.

8.1 NEXT RESEARCH STEPS: THE PIPELINE

- Publication of the results of the first section of this thesis: the statistical analysis of the currently available empirical datasets.

- Extension of the empirical datasets analysis to other international longitudinal panels and exploration of their distribution patterns when they are released.

- Parametrization and calibration of our model with the datasets already available (USA, Australia, and Sweden). Model analysis. Publication of the model and its analysis.

- Development of the underlying theoretical framework of this PhD research based in Human Behavioural Ecology (Aldrich, 2011; Davies, Krebs and West, 2012; Roundy, Bradshaw and Brockman, 2018). Entrepreneurship as a complex “human foraging”, and the relationship with non-human “entrepreneurial” behaviour (especially in the family *Hominidae* (- the great apes -).

8.2 A NEXT WORKING PAPER: ENTREPRENEURSHIP AS A FORM OF "COMPLEX HUMAN FORAGING" AND ITS ANTECENDENTS IN THE *HOMINIDAE* FAMILY

INTRODUCTION

This new line of my research is still in an infant stage: I have gathered the main bibliography and I am now in the literature review process, framing the main themes of this paper. This working paper may have a title similar to:

> "*Entrepreneurship as a Complex Human Foraging: A Human Behavioural Ecology Perspective to Human Entrepreneurship*".

The main thesis is that entrepreneurship may be analysed as a complex type of "complex human foraging". Thus, the encounter of the entrepreneur and the opportunity and the subsequent entrepreneurial tasks are part of this "complex foraging". This line of research will address the nascent entrepreneurial processes from the Human Behavioural Ecology perspective -Biological Anthropology- as theoretical framework in order to explain heavy tail distributions patterns observed worldwide in entrepreneurial longitudinal studies as we analysed in the first section of this document.

This approach will also provide the main concepts needed to explore and justify the use of agent-based modelling techniques in nascent entrepreneurship from a biological/ecological perspective. In ecology, agent-based modelling has served as a standard tool to study animal foraging in different ecosystems (Dumont and Hill, 2004; McLane et al., 2011), or behaviour pattern in primates (Hemelrijk, 2002; Bryson, Ando and

Lehmann, 2007). Therefore, agent-based modelling can be also a useful method to address entrepreneurship as complex human foraging.

THE BEHAVIOURAL ECOLOGY APPROACH TO ENTREPRENEURSHIP: ANTECEDENTS.

Behavioural Ecology is a branch of ecology stablished in the last 40 years based on concepts of ethology (Tinbergen, 1963), population genetics and ecology that research how animal behaviour adapts to the physical and social environment of individuals. It tries "*to understand how animal behaviour evolves in relation to the different ecological conditions"* (Davies, Krebs and West, 2012, p. 22). The basic assumption is that individuals develop a set of strategies of behaviour that increases their fitness in a specific context of ecological and social conditions. Behavioural patterns have evolved depending on the physical and social conditions in which animals have to survive (natural selection). Behavioural Ecology adopts different methods and tools from genetics, bioinformatics, developmental biology, physiology, primatology, neuroethology, etc. (Hager and Gini, 2012).

Human Behavioural Ecology (HBE) applies the evolutionary approach by natural selection of Behavioural Ecology to the study of human behaviour. This field has noticed a great development over the last 30 years, and it is closely related to disciplines such as "evolutionary anthropology", "human evolutionary ecology", "evolutionary biological anthropology", "human ethology", "socio-ecology", "biosocial (or biocultural) anthropology" or "sociobiology" (Borgerhoff Mulder and Schacht, 2012).

The underlying premise of Human Behavioural Ecology is the rejection of the need of different explanatory approaches for the study of human behaviour as opposed to that of any other animal. It does not imply that humans do not have distinctive cognitive and behavioural mechanisms

-- because they do -- but rather that the Behavioural Ecology scientific methodology for explaining behaviour in the animal realm remains similar to the one used for the human species behaviour, that is, to explore fitness cost and benefits given a specific ecological context, to make predictions based on fitness maximization, and test them empirically (Nettle, Gibson, Lawson, Sear, 2013, p. 1032). On the other hand, this approach is not very different than the ones used in microeconomic models also based on maximization. In fact, the current trend in Human Behavioural Ecology is to build bridges with social sciences and to introduce the adaptive evolutionary perspective in the social science literature corpus. Human Behavioural Ecology - with its broad scope and general empirical principles - claims to have the potential of being a common ground across social scientists in order to address the fragmentation of the study of human behaviour into many disciplinary areas (Nettle et al., 2013, p. 1036-7; Gibson and Lawson, 2015).

Human Behavioural Ecology belongs, therefore, to the evolutionary perspective on the study of the set of behaviours of the *Hominidae*. The evolutionary approach in the study of entrepreneurship – and in organizational studies, in general -- has been widespread in the last years, especially under the influence of the works of Howard Aldrich (Shane, 2004; Aldrich and Ruef, 2006; Aldrich, 2011). Aldrich stablished the evolutionary framework for studying entrepreneurship already in his book *Organizations and Environment* (Aldrich, 1979) but it was in his book *Organizations Evolving* (Aldrich, 1999; Aldrich and Ruef, 2nd ed., 2006) where he set the itinerary to consolidate entrepreneurship as an evolutionary field systematically (Shane, 2004). In similar way that in current evolutionary biology, the concepts of ecosystem and population have become paramount (Ridley, 2004, p. 2-3), there been also an analogous increase in the use of these evolutionary concepts such as ecosystems or populations in the study of organizations and entrepreneurship (Craig, 2013; Thomas and Autio, 2014; Roundy, Bradshaw and Brockman, 2018).

However, Aldrich's initial evolutionary approach in *Organizations Evolving* is still too "sociological" from a natural science perspective (Aldrich & Ruef, 2006; Aldrich, 2011). It does not really "integrate" the biological nature of human organizations. Evolutionary thinking is applied to organizations without effective acknowledgement and further implementations of the biological substratum of human populations. It is still a "*metaphorical"* evolutionary approach (Breslin, 2008, p. 402).

In 2004, McKelvey challenged the evolutionary research on entrepreneurship because, in his opinion, was too biased toward Darwinian determinism. He also proposed agent-based modelling to complement the evolutionary perspective with the complexity theory paradigm (McKelvey, 2004), which is a more versatile approach able to examine the creation of pattern without imposing the limitations of the Darwinian theory. However, in the last 20 years evolutionary theory in biology has undergone major conceptual changes. New discoveries in population genetics and molecular biology, have lead the field towards a new theoretical framework called "the Extended Evolutionary Synthesis" (EES) that considers non-genetic inheritance modes, such as epigenetics, parental effects, ecological inheritance, cultural inheritance, and evolvability (Laland, et al., 2015). Therefore, the current evolutionary theory is much less deterministic nowadays than it was in 2004, and complexity theory has also influenced strongly the post-neo-Darwinism (Weber, 2011).

If Aldrich – -as sociologist-- would talk of "organizational populations" of humans, this working paper will extent the "ecosystem" and "population" metaphor to the extreme: the "organizational populations" and "ecosystems" of an animal belonging to the genus *Homo* – us -- with the theoretical framework and methodological tools of Human Behavioural Ecology. Aldrich himself and many others were aware of the need of this step further, but the fragmentation between natural and social sciences has

delayed this interdisciplinary fertilization (Aldrich et al., 2008; Liguori et al., 2018).

ENTREPRENEURSHIP AS A "COMPLEX HUMAN FORAGING".

"Foraging" or "Optimal Foraging Theory" in one of the main concepts of Behavioural Ecology. Foraging theory is the study of the processes associated with resource acquisition. It studies the foraging behaviour in relation to the environment where the animal lives. Behavioural ecology mostly uses models based on optimization – or maximization – to understand foraging, that is, foraging theory analyse the set of behaviours in term of optimizing the payoff from foraging decisions – including optimization through game theory models - (Stephens and Krebs, 1986; Stephens, Brown and Ydenberg, 2007).

There are several factors that influence greatly the ability to forage and acquire profitable resources, such as learning, genetics or the presence of predators. Learning, for example, is a major factor in non-human primates, where the youngest learn by watching other group members forage and by copying their behaviours (Rapaport and Brown, 2008). There are also several types of optimal foraging depending on the different foraging situations. Optimal theory models generally have these three main components (Stephens, Brown and Ydenberg, 2007): a) **currency**, as an objective function, to be maximized (energy over time, etc.); b) the **set of behavioural choices** that the animal can control or the decisions that the animal exhibits; and c) the **animal's behavioural constraints**: such as genetics, physiology, neurology, morphology, etc. (Stephens, Brown and Ydenberg, 2007).

Please notice that not all human foraging is a "complex foraging". The initial applications of optimal foraging theory (OFT) in humans were in

the most ancient and “simpler” human foraging, closed related to those of the primates: the **foraging (hunter-gatherer) subsistence behaviour** (Winterhalder, 1981; Raichlen et al., 2014). In the last decades, Human Behavioural Ecology has also explored more recent human foraging behaviour, for example, the study of the emergence of the adoption of agriculture due to lower foraging encounter rates with higher-ranked food items, probably resulted from the late Pleistocene climatic change, in which human population increased (Richerson et al., 2001). Optimal Foraging Theory has also been extended to study patterns in modern fisheries and livestock domestication, converging closely with microeconomics models (Tucker, 2007).

This paper will introduce the concepts of “proto-entrepreneurial activities” and seminal organizations of individuals (“proto-ventures”) in order to address some “entrepreneurial-like” set of behaviours observed in non-human primates (Alcock, 2013). We will also explore the field of primatology to show how indeed some forms of “primitive entrepreneurship” can be found in non-human primates (and in other social species), based on the works of de Waal and Tyack (2003). We will define “proto-entrepreneurship” as these set of behaviours in non-human species in order to remark the differences with complex human entrepreneurship, entering therefore into Comparative Psychology and Ethology, in what we have called “Comparative Entrepreneurship” (differences between non-human and human entrepreneurship).

Aims of This Research

This paper will propose to root entrepreneurship more deeply in the biological foundations of human behaviour. Entrepreneurship, thus, may be considered an adaptive set of behaviours for survival and human development. As such, it can be studied from the human behavioural ecology perspective. Can we address the implications of entrepreneurship

as evolutionary, adaptive set of human behaviours? Is human entrepreneurship a form of “complex” foraging, a more sophisticated and evolved form of pre-human foraging? Here, the entrepreneurial activities are regarded as an adaptive set of behaviours to obtain resources that have evolved from pre-human foraging to a complex form of human foraging. Is possible to use the Behavioural Ecology methodology and tools to explore more empirically data and entrepreneurial behavioural models?

9. BIBLIOGRAPHY

Acs, Z. J. and Audretsch, D. B., eds., 2010. *Handbook of Entrepreneurship Research: An Interdisciplinary Survey and Introduction*. Second Edition. New York; London: Springer.

Adner, R., Polos, L., Ryall, M. and Sorenson, O., 2009. The Case for Formal Theory. *Academy of Management Review*, 34(2), 201-208.

Aitchison J, and Brown J.A.C., 1957. *The Log-normal Distribution with special reference to its uses in economi (sic)*. Cambridge (UK): Cambridge University Press. Reprinted 1963.

Alcock, Jon. 2013. *Animal behavior: An evolutionary approach.* New York: Sinauer.

Aldrich, H.E., 1979. *Organizations and environments.* Englewood Cliffs; London: Prentice-Hall.

Aldrich, H.E., 1999. *Organizations evolving*, London: Sage.

Aldrich, H.E., 2001. “Who wants to be an evolutionary theorist?”, Journal of Management Inquiry, Vol. 10 No. 2, pp. 115-27.

Aldrich, H.E. and Ruef, M. 2006. *Organizations Evolving*, 2nd ed., London: Sage Publications.

Aldrich, H.E., Hodgson, G.M., Hull, D.L., Knudsen, T., Mokyr, J. & Vanberg, V.J. 2008, "In defence of generalized Darwinism", *Journal of Evolutionary Economics,* vol. 18, no. 5, pp. 577-596.

Aldrich, H.E., 2011. *An Evolutionary Approach to Entrepreneurship: Selected Essays by Howard E. Aldrich.* Cheltenham: Edward Elgar.

Alstott, J., Bullmore, E. and Plenz, D., 2014. Powerlaw: a Python package for analysis of heavy-tailed distributions, *PloS One*, vol. 9, no. 1, p. e85777.

Amitrano, D., 2012. Variability in the power-law distributions of rupture events. *The European Physical Journal Special Topics*, 205, pp. 199–215.

Amaral, L. A. N., Scala, A., Barthélémy, M., & Stanley, H. E., 2000. Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97, pp. 11149–11152.

Amorós, J.E., and Bosma, N., 2014. *Global Entrepreneurship Monitor 2013. Global Report: Fifteen Years of Assessing Entrepreneurship Across the Globe.* Global Entrepreneurship Research Association (GERA), London: London Business School.

http://www.gemconsortium.org/docs/download/3106

Allen, P., Maguire, S., McKelvey, B., eds., 2011. *The SAGE Handbook of Complexity and Management.* London: Sage.

Alvarez, S. A. and Barney, J. B., 2007. Discovery and creation: alternative theories of entrepreneurial action. *Strategical Entrepreneurship Journal*, 1: 11–26.

Anderson, P., 1999. Complexity Theory and Organization Science. *Organization Science*, Vol. 10, No. 3, Special Issue: Application of Complexity Theory to Organization Science (May - Jun., 1999), pp. 216-232.

Andriani, P. and McKelvey, B., 2007. Beyond Gaussian averages: redirecting international business and management research toward extreme events and power laws. *Journal of International Business Studies*, Vol. 38 No. 7, pp. 1212-30.

Andriani, P. and McKelvey, B., 2009. From Gaussian to Paretian Thinking: Causes and Implications of Power Laws in Organizations. *Organization Science,* Vol. 20, (No. 6), pp. 1053-1071.

Antoniou, I., Ivanov, V.V., Ivanov, V.V. & Zrelov, P.V., 2004. On the log-normal distribution of stock market dat., *Physica A: Statistical Mechanics and its Applications*, vol. 331, no. 3, pp. 617-638.

Aoki, M., and Yoshikawa, H., 2006. Stock prices and the real economy: Power law versus exponential distributions. *Journal of Economic Interaction and Coordination*, 1, pp. 45–73.

Arenius, P., Engel, Y. and Klyver, K., 2017. No particular action needed? A necessary condition analysis of gestation activities and firm emergence. *Journal of Business Venturing Insights,* vol. 8, pp. 87-92.

Arthur W.B., Durlauf S.N. and Lane D.A., eds., 1997. *The Economy as an Evolving Complex System II*, SFI Studies in the Sciences of Complexity. Reading, MA: Addison-Wesley.

Arshed, N., Carter, S., Mason, C., 2014. The ineffectiveness of entrepreneurship policy: Is policy to blame? *Small Bus Econ*, 43, pp. 639–659.

Augusiak J., Van den Brink P.J., Grimm V., 2014. Merging validation and evaluation of ecological models to 'evaludation': a review of terminology and a practical approach. *Ecological Modelling,* 280: 117-128.

Axtell, R., Axelrod, J., Epstein, M., and Cohen, M., 1996. Aligning simulation models: A case study and results. *Computational & Mathematical Organization Theory, 1*(2), 123-141.

Axtell, R., 1999*. The emergence of firms in a population of agents: local increasing returns, unstable Nash equilibria, and power law size distributions.* Santa Fe Institute Working Papers, 99-03-019. Santa Fe (NM): Santa Fe Institute.

Axtell, R. L., 2001. Zipf distribution of U.S. firm sizes. *Science*, 293, pp. 1818-1820.

Axelrod, R., 1997. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton, NJ: Princeton University Press.

Axelrod, R., 1997. Advancing the art of simulation in the social sciences. *Complexity* 3, pp. 193–199.

Axelrod, R. and Tesfatsion, L., 2006. A Guide for Newcomers to Agent-Based Modelling in the Social Sciences Appendix A. In: Leigh Tesfatsion and Kenneth L. Judd (Eds.). *Handbook of Computational Economics*, *Vol. 2:*

Agent-Based Computational Economics. Elsevier B.V.: Amsterdam (pp. 1647-1659).

Axelrod R and Tesfatsion L., 2018. On-line guide for newcomers to agent-based modeling in the social sciences. Available from:

http://www2.econ.iastate.edu/tesfatsi/abmread.htm [accessed 15 August 2018].

Bak, P., Tang, C., and Wiesenfeld, K., 1987. Self-organized criticality: An explanation of the 1/f noise. *Phys. Rev. Lett.* 59, pp. 381–384.

Bak, P., Tang, C., 1989. Earthquakes as a self-organized critical phenomenon. *Journal of Geophysical Research,* 94, 15635–15637.

Bak, P., and Sneppen, K., 1993. Punctuated equilibrium and criticality in a simple model of evolution*. Phys. Rev. Lett.* 74, pp. 4083–4086.

Bak, P., 1996. *How Nature Works: The Science of Self- Organized Criticality.* New York, NY: Copernicus.

Balasooriya, U., and Abeysinghe, T., 1994. Selecting between gamma and Weibull distributions: An approach based on predictions of order statistics. *Journal of Applied Statistics*, 21, pp. 17–27.

Balke, T., and Gilbert, N., 2014. How do agents make decisions? A survey. *Journal of Artificial Societies and Social Simulation,* 17(4): 13.

http://jasss.soc.surrey.ac.uk/17/4/13.html

Banerjee, A., and Yakovenko, V. M., 2010. Universal patterns of inequality. *New Journal of Physics*, 12, p. 075032.

Barabási, A.-L., and Albert, R., 1999. Emergence of scaling in random networks. *Science* 286 (5439), pp. 509–512.

Barabási, A. L., and Bonabeau, E., 2003. Scale-free networks. *Scientific American*, 288, pp. 60–69.

Barabási, A. L. 2012. Network science: Luck or reason. *Nature*, 489, pp. 507–508.

Baron, R.A., 2007. Behavioral and cognitive factors in entrepreneurship: entrepreneurs as the active element in new venture creation, *Strategic Entrepreneurship Journal*, Vol. 1 Nos 1-2, pp. 167-82.

Barro, Robert J., and Tao Jin. 2011. On the size distribution of macroeconomic disasters. *Econometrica* 79(5), pp. 1567-1589.

Baumol, W. J., 1996. Entrepreneurship: Productive, Unproductive, and Destructive. *Journal of Political Economy*, 98 (5), pp. 3-22.

Baumol, W. J., and Schilling, M. A., 2008. Entrepreneurship. In: Steven N. Durlauf and Lawrence E. Blume, Eds. *The New Palgrave Dictionary of Economics*. Second Edition. Basingstoke: Palgrave Macmillan.

Bee, M., Riccaboni, M. and Schiavo, S., 2017. Where Gibrat meets Zipf: Scale and scope of French firms. *Physica A: Statistical Mechanics and its Applications,* vol. 481, pp. 265-275.

Bhawe, N., Rawhouser, H. and Pollack, J.M., 2016. Horse and cart: the role of resource acquisition order in new ventures. *Journal of Business Venturing Insights*, Vol. 6, pp. 7-13.

Bird, B., 1988. Implementing entrepreneurial ideas: The case for intention. *Academy of Management Review*, 13, pp. 442-453.

Bird, B., 1992. The operations of intentions in time: The emergence of new ventures. *Entrepreneurship: Theory and Practice*, Fall 92, Vol. 17 Issue 1, pp. 11-20.

Boisot, M., and McKelvey, B., 2011. Connectivity, extremes, and adaptation: A power-law perspective of organizational effectiveness. *Journal of Management Inquiry*, 20, pp. 119–133.

Bonabeau, E., 2002a. Predicting the unpredictable. *Harvard Business Review*, Vol. 80 No. 3, pp. 109-16.

Bonabeau, E., 2002b. Agent based modeling: methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, Vol. 99 No. 3, pp. 7280-7.

Borgerhoff Mulder, M, and Schacht, R. 2012. Human Behavioural Ecology. In: *Encyclopaedia of Life Sciences*. Chichester: John Wiley & Sons Ltd,. http://www.els.net

Bowles, S., 2006. *Microeconomics: Behavior, Institutions, and Evolution*. Princeton, NJ: Princeton University Press.

Breig, R., Coblenz, M. and Pelz, M., 2018. Enhancing simulation-based theory development in entrepreneurship through statistical validation. *Journal of Business Venturing Insights*, vol. 9, pp. 53-59.

Breslin, D. 2008. A review of the evolutionary approach to the study of entrepreneurship, *International Journal of Management Reviews,* vol. 10, no. 4, pp. 399-423.

Brown, D.G., Page, S., Riolo, R., Zellner, M. & Rand, W., 2005. Path dependence and the validation of agent-based spatial models of land use. *International Journal of Geographical Information Science,* vol. 19, no. 2, pp. 153-174.

Brown, J.H. & West, G.B., 2000. *Scaling in Biology.* Oxford: Oxford University Press.

Bryson, J. J., Ando, Y., and Lehmann, H. (2007). Agent-based models as scientific methodology: A case study analysing primate social behaviour. *Philosophical Transactions of the Royal Society, B - Biology*, 362(1485):1685–1698.

Brzezinski, M., 2014, "Do wealth distributions follow power laws? Evidence from 'rich lists'", *Physica A: Statistical Mechanics and its Applications,* vol. 406, pp. 155-162.

Buldyrev, S. V., Pammolli, F., Riccaboni, M. and Stanley, H. E. 2013. The Rise and Fall of the Firm. Book Draft. Preprint.

http://polymer.bu.edu/hes/bprs23dec13full.pdf

Cabral, L. M., and Mata, J., 2003. On the evolution of the firm size distribution: Facts and theory. *The American Economic Review*, Vol. 93, No. 4 (Sep.), pp. 1075-1090.

Carter, N., Gartner, W., Reynolds, P., 1996. Exploring start-up event sequences. *Journal of Business Venturing*, 11, 151–166.

Caves, R. E., 1998. Industrial organization and new findings on the turnover and mobility of firms. *Journal of Economic Literature*, Vol. 36, No. 4 (Dec.), pp. 1947-1982.

Cederman, L.-E., 2005. Computational Models of Social Forms: Advancing Generative Process Theory. *American Journal of Sociology*, Vol. 110, No. 4, January, pp. 864-893.

Cefis E, Ciccarelli M., Orsenigo L., 2007. Testing Gibrat's legacy: a Bayesian approach to study the growth of firms. *Structural Change and Economic Dynamics,* 18(3): pp. 348–369.

Cefis, E., Marsili, O., Schenk, H., 2009. The effects of mergers and acquisitions on the firm size distribution. *Journal of Evolutionary Economics*, February, Volume 19, Issue 1, pp 1-20.

Champernowne, D., 1953. A model of income distribution. *Economic Journal*, 83, 318–51.

Cheng, Y., and van de Ven, A. H., 1996. Learning the Innovation Journey: Order out of Chaos?. *Organization Science*, 7(6), pp. 593-614.

Chiles, T. H., Tuggle, C. S., McMullen, J. S., Bierman, L. and Greening, D. W., 2010. Dynamic creation: extending the radical Austrian approach to entrepreneurship. *Organization Studies*, 31, pp. 7–46.

Cirillo, P., 2013. Are your data really Pareto distributed? *Physica A: Statistical Mechanics and its Applications,* vol. 392, no. 23, pp. 5947-5962.

Clauset, A., Shalizi, C.R., and Newman, M. E. J., 2009. Power-Law Distributions in Empirical Data. *Society for Industrial and Applied Mathematics Review*, 51(4), pp. 661–703.

Coad, A., 2009. *The Growth Of Firms. A Survey of Theories and Empirical Evidence.* Cheltenham UK: Edward Elgar Publishing.

Coad, A., Frankish, J., Roberts, R.G., Storey, D.J., 2013. Growth paths and survival chances: An application of Gambler's Ruin theory. *Journal of Business Venturing*, Volume 28, Issue 5, September 2013, pp. 615-632.

Cook, W., and Ormerod, P., 2003. Power law distribution of the frequency of demises of US firms, *Physica A: Statistical Mechanics and its Applications*, Volume 324, Issues 1–2, 1 June, pp. 207-212.

Cooper, A.C., 1993. Challenges in predicting new firm performance. *Journal of Business Venturing,* 8, 241–253.

Corbett, A.C., 2007. Learning asymmetries and the discovery of entrepreneurial opportunities. *Journal of Business Venturing*, Vol. 22 No. 1, pp. 97-118.

Coviello, N.E. and Jones, M.V., 2004. Methodological issues in international entrepreneurship research. *Journal of Business Venturing*, Vol. 19 No. 4, pp. 485-508.

Craig, J.B. 2013. An Evolutionary Approach to Entrepresneurship: Selected Essays, *Academy Of Management Learning & Education*, 12, 1, pp. 145-146.

Crawford, G. C. and McKelvey, B., 2012. Strategic Implications of Power-Law Distributions in the Creation and Emergence of New Ventures: Power-Law Analyses in three Panel Studies. *Frontiers of Entrepreneurship Research*, Vol. 32, Issue 12, Article 1, p. 1-15.

Crawford, G. C., 2012a. Disobeying Power-Laws: Perils for Theory and Method. *Journal of Organization Design* 1(2), pp. 75-81.

Crawford, G. C., 2012b. Toward a Scale-Free Theory of New Venture Performance: A Complexity Science Approach Through The Lens of Regulatory Focus Theory (Summary). *Frontiers of Entrepreneurship Research*: Vol. 32: Iss. 5, Article 6.

Crawford, C. and Lichtenstein, B., 2013. Is There A Single Driver of Entrepreneurship? A Power-Law of Organizational Emergence and Growth. *Academy of Management Annual Conference Best-Paper Proceedings.* August, pp. 776-781.

Crawford, G. C., McKelvey, B., Lichtenstein, B. B., 2014. The empirical reality of entrepreneurship: How power law distributed outcomes call for new theory and method. *Journal of Business Venturing Insights* 1 (1–2), pp. 3–7.

Crawford, G. C., Aguinis, H., Lichtenstein, B., Davidsson, P., and McKelvey, B., 2015. Power law distributions in entrepreneurship: Implications for theory and research. *Journal of Business Venturing*, 30, pp. 696–713.

Cullen, A. and Frey, H. (1999). *Probabilistic Techniques in Exposure Assessment.* 1st edition. s.l.: Plenum Publishing Co.

Curtis, R. and Reynolds, Paul D., 2007. Second Panel Study of Entrepreneurial Dynamics: Research Opportunities with PSED II. Professional Development Workshop, *Academy of Management Annual Meeting* (3 August), Philadelphia, PA

http://www.psed.isr.umich.edu/psed/download_document/14

Curtin, R., 2012. *Panel Study of Entrepreneurial Dynamics II: Codebook*. Ann Arbor, MI: University of Michigan.

Dasi, G., Marsili, O., Orsenigo, L. and Salvatore, R., 1995. Learning, market selection and the evolution of industrial structures, *Small Business Economics*, 7, pp. 411–436.

Davidsson, P., and Wiklund, J., 2001. Levels of Analysis in Entrepreneurship Research: Current Research Practice and Suggestions for the Future. *Entrepreneurship: Theory & Practice*, 25(4), pp. 81-99.

Davidsson, P., 2003. The Domain of Entrepreneurship Research: Some suggestions. In: J. Katz & S. Shepherd, Eds., *Advances in Entrepreneurship, Firm Emergence and Growth,* Vol.6, pp. 315-372. Oxford, UK: Elsevier/JAI Press.

Davidsson, P. and Honig, B., 2003. The role of social and human capital among nascent entrepreneurs. *Journal Business Venturing*, 18, pp. 301–31.

Davidsson, P., 2006. Nascent entrepreneurship: empirical studies and developments. *Foundations and Trends in Entrepreneurship*, 2, pp. 1–76.

Davidsson, P., and Steffens, P., 2011. Comprehensive Australian Study of Entrepreneurial Emergence (CAUSEE): Project Presentation and Early Results. In: P. D. Reynolds & R. T. Curtin (Eds.), *New Business Creation* New York: Springer, pp. 27-51.

Davidsson, P., Steffens, P. and Gordon, S.R., 2011. Comprehensive Australian Study of Entrepreneurial Emergence (CAUSEE): Design, Data Collection and Sample Description. In: Hindle, K & Klyver, K., *Handbook of New Venture Creation Research.* Cheltenham: Elgar, pp. 216-250.

Davidsson, P. and Gordon, S.R., 2012. Panel studies of new venture creation: a methods-focused review and suggestions for future research. *Small Business Economics*, Vol. 39 No. 4, pp. 853-876.

Davis, J.P., Eisenhardt, K.M. and Bingham, C.B., 2007. Developing Theory Through Simulation Methods. *Academy of Management Review,* 32(2), pp. 480-499.

Davis, J.P., Eisenhardt, K.M. and Bingham, C.B., 2009. Optimal structure, market dynamism, and the strategy of simple rules. *Administrative Science Quarterly*, Vol. 54, No. 3, pp. 413-52.

Davies, N.B., Krebs, J.R. and West, S.A., 2012, *An introduction to behavioural ecology,* 4th ed, Oxford: Wiley-Blackwell.

D'Agostino, R. and Stephens, M., 1986. *Goodness-of-Fit Techniques*. 1st edition. s.l.: Dekker.

de Holan, P.M., 2014. It's All in Your Head: Why We Need Neuroentrepreneurship. *Journal of Management Inquiry,* vol. 23, no. 1, pp. 93-97.

De Waal, F.B.M and Tyack, P.L. 2003. *Animal Social Complexity. Intelligence, Culture, and Individualized Societies*. Cambridge, MA: Harvard University Press.

Delignette-Muller, M., Pouillot, R., Denis, J., and Dutang, C., 2014. **fitdistrplus**: Help to Fit of a Parametric Distribution to Non-Censored or Censored Data. R package version 1.0-2.

Delignette-Muller, M.L. and Dutang, C., 2015. "fitdistrplus: An R Package for Fitting Distributions", *Journal of Statistical Software,* vol. 64, no. 4.

Delignette-Muller M.L., Pouillot R., Denis J.B., and Dutang, C., 2015. fittdistrplus: Help to Fit of a Parametric Distribution to Non-Censored or Censored Data. R package version 1.0-4,

Available from: http://CRAN.R-project.org/package=fitdistrplus [Accessed 1 September 2018]

Delmar, F. and Shane, S., 2003. Does business planning facilitate the development of new ventures? *Strategic Management Journal*, 24, pp. 1165–85.

Delmar, F. and Shane, S., 2004. Legitimizing first: organizing activities and the survival of new ventures. *Journal of Business Venturing*, 19, pp. 385–410.

Dew, N., Read, S., Sarasvathy, S.D. and Wiltbank, R., 2008. Outlines of a behavioral theory of the entrepreneurial firm. *Journal of Economic Behavior and Organization,* vol. 66, no. 1, pp. 37-59.

Dimov, D., 2010. Nascent Entrepreneurs and Venture Emergence: Opportunity Confidence, Human Capital, and Early Planning. *Journal of Management Studies*, 2010, Vol.47 (6), pp.1123-1153.

Dimov, D., 2011. Grappling with the Unbearable Elusiveness of Entrepreneurial Opportunities. *Entrepreneurship: Theory and Practice,* vol. 35, no. 1, pp. 57-81.

Di Guilmi, C., Gallegati, M., Ormerod, P., 2004. Scaling invariant distributions of firms' exit in OECD countries. *Physica A: Statistical Mechanics and its Applications*, Volume 334, Issues 1–2, 1 March 2004, pp. 267-273.

Drăgulescu, A. and Yakovenko, V.M., 2001a. Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. *Physica A,* 299, pp. 213–221.

Drăgulescu, A. and Yakovenko, V.M., 2001b. Evidence for the exponential distribution of income in the USA. *The European Physical Journal B*, vol. 20, no. 4, pp. 585-589.

Drossel, B, and Schwabl, F., 1992. Self-organized critical forest fire model. *Phys. Rev. Lett.,* 69, pp. 1629–1632.

Dumont, B., and Hill, D. R. 2004. Spatially explicit models of group foraging by herbivores: what can Agent-Based Models offer?. *Animal Research*, *53*(5), 419-428.

Durlauf, S. N., 2005. Complexity and Empirical Economics. *The Economic Journal*. Vol. 115, No. 504, Features (Jun.,), pp. F225-F243.

Durlauf, S. N. 2012. Complexity, economics, and public policy. *Politics, Philosophy & Economics*, February, 11(1), pp. 45-75.

Easley, D. and Kleinberg, J., 2010. *Networks, Crowds, and Markets.* Cambridge: Cambridge University Press.

Epstein, B., 1947. The mathematical description of certain breakage mechanisms leading to the logarithmic-normal distribution, *Journal of the Franklin Institute*, Volume 244, Issue 6, 1947, pp. 471-477.

Epstein, J.M., Axtell, R., 1996. *Growing Artificial Societies: Social Science from the Bottom Up.* Cambridge, MA: MIT Press.

Epstein, J.M., 1999. Agent-based computational models and generative social science. *Complexity*, 4 (5), pp. 41–57.

Epstein, J.M., 2006a. Chapter 34: Remarks On The Foundations of Agent-Based Generative Social Science. In: Leigh Tesfatsion and Kenneth L. Judd

(Eds.). *Handbook of Computational Economics*, Volume 2. Elsevier B.V.: Amsterdam.

Epstein, J.M. 2006b. *Generative social science: Studies in agent-based computational modeling.* Princeton: Princeton University Press.

Epstein, J.M., 2008. 'Why Model*? Journal of Artificial Societies and Social Simulation*, 11, 4.

Epstein J.M., 2014. Agent_Zero: Toward Neurocognitive Generative Social Science. Princeton, NJ: Princeton University Press.

Evans, D. S., 1987. The Relationship between Firm Growth, Size, and Age: Estimates for 100 Manufacturing Industries. *Journal of Industrial Economics*, June, 35(4), pp. 567-81.

Farmer, J. D. and Foley, D. 2009. The economy needs agent-based modelling. *Nature,* 460, pp. 685-6.

Fioretti, G., 2013. Agent-based simulation models in organization science. *Organizational Research Methods*, Vol. 16, No. 2, pp. 227-242.

Frankish, J. S., Roberts, R. G., Coad, A., Spearsz, T.C., and Storey, D. J., 2013. Do entrepreneurs really learn? Or do they just tell us that they do? *Industrial and Corporate Change*, Volume 22, Number 1, pp. 73–106.

Freeman, J. and Hannan, M.T., 1983. Niche width and the dynamics of organizational populations. *American Journal of Sociology*, Vol. 88 No. 6, pp. 1116-45.

Fujiwara, Y., Di Guilmi, C., Aoyama, H., Gallegati, M. and Souma, W., 2004. Do Pareto–Zipf and Gibrat laws hold true? An analysis with European firms. *Physica A,* 335, pp. 197–216.

Fujiwara, Y., 2004. Zipf Law in Firms Bankruptcy. *Physica A,* 337, 219–30.

Gabaix, X., 1999. Zipf's law for cities: an explanation. *Quarterly Journal of Economics* 114, pp. 739–67.

Gabaix, X., Gopikrishnan, P., Plerou, V. and Stanley, H.E. 2003. A theory of power law distributions in financial market fluctuations. *Nature,* 423, pp. 267–30.

Gabaix, X. and Ioannides, Y. 2004. The evolution of the city size distributions. In *Handbook of Regional and Urban Economics*, vol. 4, ed. V. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Gabaix, X. and Ibragimov, R. 2006. Log (Rank-1/2): a simple way to improve the OLS estimation of tail exponents. Working paper, Harvard University.

Gabaix, X., 2008. Power Laws. In: Steven N. Durlauf and Lawrence E. Blume, Eds., *The New Palgrave Dictionary of Economics.* Second Edition. Basingstoke: Palgrave Macmillan.

Gabaix, X. and Landier, A. 2008. Why has CEO pay increased so much? *Quarterly Journal of Economics,* 123(1), pp. 49—100.

Gabaix, X., 2009. Power laws in economics and finance. *Annual Review of Economics,* 1, pp. 255–293.

Gabaix, X., 2014. *Power Laws in Economics: An Introduction.* Working paper at http://pages.stern.nyu.edu/~xgabaix/

Gaddum, J.H., 1945. Log normal distributions. *Nature*, 156, (22 December), pp. 746–747.

Gaffeo E., Di Guilmi C., Gallegati M., Russo A., 2012. On the mean/variance relationship of the firm size distribution: evidence and some theory. *Ecological Complexity*, Volume 11, September 2012, pp. 109-117.

Galton, F., 1879. The geometric mean, in vital and social statistics. *Proceedings of the Royal Society*, 29, pp. 365–367.

Ganco, M. and Agarwal, R., 2009. Performance differentials between diversifying entrants and entrepreneurial start-ups: A Complexity approach. *Academy of Management Review*, 34, pp. 228-253.

Gartner, W.B., 1985. A conceptual framework for describing the phenomenon of new venture creation. *Academy of Management Review*, 10(4), pp. 696-706.

Gartner, W.B., Bird, B.J., and Starr, J.A., 1992. Acting as if: Differentiating entrepreneurial from organizational behavior. *Entrepreneurship: Theory and Practice,* 16(3), pp. 13-31.

Gartner, W.B. and Shane, S.A., 1995. Measuring entrepreneurship over time. *Journal of Business Venturing*, Vol. 10, No. 4, pp. 283-331.

Gartner, W. B., Shaver, K. G. and Liao, J., 2008. Opportunities as attributions: Categorizing strategic issues from an attributional perspective. *Strategical Entrepreneurship Journal*, 2, pp. 301–315.

Gartner, W. B., Carter, N. M. and Reynolds, P. D., 2010. Entrepreneurial Behaviour: Firm Organizing Processes. In: Acs, Z. J. and Audretsch, D. B. (Eds), *Handbook of Entrepreneurship Research: An Interdisciplinary Survey and Introduction.* Second Edition. New York; London: Springer, pp. 99-128.

Gatewood, E., Shaver, K., Gartner, B., 1995. A longitudinal study of cognitive factors influencing start-up behaviors and success at venture creation. *Journal of Business Venturing,* 10 (5), pp. 371–391.

Gibrat R. 1931. Les Inégalités Economiques. Paris: Recueil Sirey.

Gibson, M.A. and Lawson, D.W., 2015. Applying evolutionary anthropology. Evolutionary anthropology, vol 24, no. 1, pp. 3-14.

Gilbert, N. and Terna, P., 2000. How to build and use agent-based models in social science. *Mind and Society*, Vol. 1 No. 1, pp. 57-72.

Gilbert, N., 2008. *Agent-Based Models.* Thousand Oaks, CA: Sage Publications.

Gillespie, C. S., 2015. Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software*, 64(2), pp. 1-16.

https://github.com/csgillespie/poweRlaw

Ginsburg. A, 2012. **plfit** 1.0.2 [Python implementation of Aaron Clauset's power-law distribution fitter]. Available from: https://pypi.python.org/pypi/plfit. [Accessed 21 January 2018].

Goldstein, J., 1999. Emergence as a Construct: History and Issues. *Emergence: Complexity and Organization,* 1 (1), pp. 49–72.

Goldstein, J., 2011. Emergence in Complex Systems. In: Allen, P., Maguire, S., McKelvey, B., eds., 2011. *The SAGE Handbook of Complexity and Management.* London: Sage, pp. 65-78.

Gonzalez-Estrada, E., and Villasenor-Alva, J.A., 2017. Tests of Fit for some Probability Distributions (Package: **goft**) Version: 1.3.4, Repository: CRAN

Gopikrishnan, P., Plerou, V., Amaral, L., Meyer, M. and Stanley, H.E., 1999. Scaling of the distribution of fluctuations of financial market indices. *Physical Review,* E 60, pp. 5305–316.

Gopikrishnan, P., Plerou, V., Gabaix, X. and Stanley, H.E., 2000. Statistical properties of share volume traded in financial markets. *Physical Review* E, 62, R4493–96.

Gordon, D. M., 1999. Interaction patterns and task allocation in ant colonies. In: C. Detrain, J M Pasteels, J L Deneubourg, eds., *Information Processing in Social Insects.* Basel, Boston, Berlin: Birkhäser Verlag: 51-67.

Gordon, D.M., 2007. Control without hierarchy. *Nature*, Vol 446, 8 March 2007, p.143ff.

Grimm, V., Railsback, S., 2005. Individual-based Modeling and Ecology. Princeton, NJ: Princeton University Press.

Granovetter, M., 1973. The strength of weak ties. *American Journal of Sociology*, Vol. 78 No. 6, pp. 1360-80.

Grimm, V. and Railsback, S.F., 2005. *Individual-Based Modeling and Ecology*, Princeton, NJ: Princeton University Press.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H.-H., Weiner, J., Wiegand, T., Donald, L. and DeAngelis, D.L., 2005. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science*, Vol. 310, No. 5750, pp. 987-91.

Grimm V, Berger U, Bastiansen F, Eliassen S, Ginot V, Giske J, Goss-Custard J, Grand T, Heinz S, Huse G, Huth A, Jepsen JU, Jørgensen C, Mooij WM, Müller B, Pe'er G, Piou C, Railsback SF, Robbins AM, Robbins MM, Rossmanith E, Rüger N, Strand E, Souissi S, Stillman RA, Vabø R, Visser U, DeAngelis D.L., 2006. A standard protocol for describing individual-based and agent-based models. *Ecological Modelling,* 198, pp.115-126.

Grimm, V., Ashauer, R., Forbes, V., Hommen, U., Preuss, T.G., Schmidt, A., van denBrink, P.J., Wogram, J., Thorbek, P., 2009. CREAM: A European project on mechanistic effect models for ecological risk assessment of chemicals. *Environmental Science and Pollution Research,* 16, pp. 614–617.

Grimm V., Berger U., DeAngelis D.L., Polhill, G., Giske, J., Railsback, S.F., 2010. The ODD protocol: a review and first update. *Ecological Modelling,* 221, pp. 2760-2768.

Grimm, V., Augusiak, J., Focks, A., Frank, B.M., Gabsi, F., Johnston, A.S.A., Liu, C., Martin, B.T., Meli, M., Radchuk, V., Thorbek, P. & Railsback, S.F., 2014. Towards better modelling and decision support: documenting model development, testing, and analysis using TRACE. *Ecological Modelling,* vol. 280, pp. 129-139.

Grimm, V. and Berger, U., 2016. Robustness analysis: Deconstructing computational models for ecological theory and applications. *Ecological Modelling,* vol. 326, pp. 162-167.

Gruenhagen, J., Davidsson, P., Gordon, S.R., Salunke, S., Senyard, J., Steffens, P., & Stuetzer, M., 2016. *Comprehensive Australian Study of Entrepreneurial Emergence (CAUSEE). Handbook & User Manual*, Version 7. Brisbane: Australian Centre for Entrepreneurship Research (ACE) at QUT Business School.

Günther, M., Stummer, C., Wakolbinger, L.M. and Wildpaner, M., 2011. An agent-based simulation approach for the new product diffusion of a novel biomass fuel. *Journal of the Operational Research Society*, Vol. 62, No. 1, pp. 12-20.

Gupta, R. D., and Kundu, D., 1999. Generalized exponential distributions. *Australian & New Zealand Journal of Statistics*, 41, pp. 173–188.

Gupta, R. D., and Kundu, D., 2001. Exponentiated exponential family: An alternative to gamma and Weibull distributions. *Biometrical Journal Biometrische Zeitschrift*, 43, pp. 117–130.

Gutenberg, B. and Richter, C.F., 1954. *Seismicity of the Earth and Associated Phenomena*. 2nd ed. Princeton: N.J.: Princeton University Press.

Hager, R. and Gini, B. 2012. Behavioural Ecology. In: *Encyclopaedia of Life Sciences*. Chichester: John Wiley & Sons Ltd. http://www.els.net

Hamill, L. and Gilbert, G.N., 2016. *Agent-based modelling in economics.* Chichester, UK: John Wiley & Sons.

Headd B., 2003. Redefining Business Success: Distinguishing Between Closure and Failure. *Small Business Economics*, August 2003, 21(1), p. 51.

Heath, B., Hill, R., Ciarallo, F., 2009. A survey of agent-based modeling practices (January 1998 to July 2008). *J. Artif. Soc. Soc. Simul.*, 12, pp. 9.

Heiner, R.A., 1988. The necessity of imperfect decisions. *Journal of Economic Behavior & Organization*, Vol. 10 No. 1, pp. 29-55.

Hemelrijk, C. 2002. Despotic Societies, Sexual Attraction and the Emergence of Male 'Tolerance': An Agent-Based Model. *Behaviour, 139*(6), 729-747.

Herdan, G., 1953. *Small-particles Statistics*. Amsterdam: Elsevier.

Haken, H., 2008. Self-organization. *Scholarpedia*, 3(8), pp. 1401.

Hall, B.H., 1987. The Relationship Between Firm Size and Firm Growth in the U.S. Manufacturing Sector. *Journal of Industrial Economics*, Vol. 35 (4), pp. 583–606.

Hannan, M.T. and Freeman, J., 1977. The population ecology of organizations. *American Journal of Sociology*, Vol. 82, No. 5, pp. 929-64.

Hart, P.E. and Prais, S.J., 1956. The Analysis of Business Concentration. *Journal of the Royal Statistical Society*, Vol. 119, pp. 150–191.

Harrison, J.R., Lin, Z., Carroll, G.R. and Carley, K.M., 2007. Simulation modeling in organizational and management research. *Academy of Management Review*, Vol. 32, No. 4, pp. 1229-45.

Hausmann, R., Hidalgo, CA et al., 2011. *The Atlas of Economic Complexity*, Cambridge MA: Puritan Press. Version online available from:

http://atlas.media.mit.edu/book/

Holland, J., 1995. *Hidden Order: How Adaptation Builds Complexity.* Addison-Wesley: Reading, MA.

Holt, R. F., Rosser, J. B., and Colander, D. 2011. The Complexity Era in Economics. *Review Of Political Economy,* 23(3), pp. 357-369.

Honig, B. and M. S., 2012. Planning and the entrepreneur: a longitudinal examination of nascent entrepreneurs in Sweden. *Journal of Small Business Management,* 50 (3), pp. 365-388.

Hopenhayn, H. A., 1992. Entry, exit, and firm dynamics in long run equilibrium. *Econometrica*, Vol. 60, No. 5 (Sep.), pp. 1127-1150.

Ijiri Y., and Simon H.A., 1964. Business firm growth and size. *The American Economic Review*, Vol. 54, No. 2, Part 1 (March), pp. 77-89.

Ijiri, Y., and Simon, H. A., 1974. Interpretations of departures from the Pareto curve firm-size distributions. *The Journal of Political Economy*, vol. 82, no. 2, pp. 315-331.

Jensen, H. J., 1998. *Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems.* Cambridge: Cambridge University Press.

Joo, H., Aguinis, H., and Bradley, K. J., 2017. Not all nonnormal distributions are created equal: Improved theoretical and measurement precision. *Journal of Applied Psychology, 102* (7), pp. 1022-1053.

Jovanovic, B., 1982. Selection and the Evolution of Industry. *Econometrica,* Vol. 50, No. 3, pp. 649-670.

Kalecki, M., 1945. On the Gibrat distribution. *Econometrica*, 13, p. 161.

Kapteyn, J.C., 1903. *Skew Frequency Curves in Biology and Statistics*. Astronomica Laboratory, Groningen: Noordhoff.

Katz, J., and Gartner, W.B., 1988. Properties of emerging organizations. *Academy of Management Review,* 13(3), pp. 429-442.

Kauffman, S.A., 1993. *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press: Oxford, UK.

Klass, O.S., Biham, O., Levy, M.,Malcai, O., and Solomon, S., 2006. The Forbes 400 and the Pareto Wealth Distribution. *Economics Letters*, 2006, 90, pp. 290–5.

Kolmogorov, AN., 1941. Über das Logarithmisch Normale Verteilungsgesetz der Dimensionen der Teilchen bei Zerstückelung, *Doklady Akademii Nauk USSR (C.R. Acad. Sci. USSR)*, 31, pp. 99–101.

.Koch, A. L., 1966. The logarithm in biology I. Mechanisms generating the log-normal distribution exactly. *Journal of Theoretical Biology*, Volume 12, Issue 2, November, pp. 276-290.

.Koch, A. L., 1969. The logarithm in biology: II. Distributions simulating the log-normal. *Journal of Theoretical Biology*, Volume 23, Issue 2, May, pp. 251-268.

Kuhn, T.S. 1996. *The structure of scientific revolutions.* 3rd edn, University of Chicago Press: London, Chicago.

Kuckertz, A., Harms, R., Semrau, T., 2014: Essential Readings in Entrepreneurship.

Available online at: https://entrepreneurship.uni-hohenheim.de/essential-readings-englisch.

[Accessed 4 November 2014]

Lamberson, P. J. and Page, S. E., 2012. *Tipping Points. Santa Fe Institute Working Papers* 2012-02-002, Santa Fe, New Mexico: Santa Fe Institute. Available at:

http://www.santafe.edu/media/workingpapers/12-02-002.pdf

[Accessed 6 October 2016]

Laland, K.N., Uller, T., Feldman, M.W., Sterelny, K., Muller, G.B., Moczek, A., Jablonka, E., Odling-Smee, J. 2015. The extended evolutionary synthesis: its structure, assumptions and predictions. *Proc. R. Soc. B* 282: 20151019.

Leavitt, K., Mitchell, T. R., and Peterson, J., 2010. Theory pruning: Strategies to reduce our dense theoretical landscape. *Organizational Research Methods*, 13, pp. 644–667.

Levy, F., 1987. Changes in the Distribution of American Family Incomes, 1947 to 1984. *Science*, Vol. 236, Issue 4804, (22 May), pp. 923-927.

Li, T. and Gustafsson, V., 2012. Nascent entrepreneurs in China: social class identity, prior experience affiliation and identification of innovative opportunity: A study based on the Chinese Panel Study of Entrepreneurial Dynamics (CPSED) project. *Chinese Management Studies*, Vol. 6 Iss: 1, pp.14 – 35

Lichtenstein, B., K. Dooley, and T. Lumpkin. 2006. Measuring emergence in the dynamics of new venture creation. *Journal of Business Venturing,* 21, pp. 153-175.

Lichtenstein, B., Carter, N., Dooley, K. and W. Gartner, W., 2007. Complexity dynamics of nascent entrepreneurship. *Journal of Business Venturing,* 22, (2), pp. 236-261.

Lichtenstein, B., 2009. Moving Far From Far-From-Equilibrium: Opportunity Tension as the Driver of Emergence. *Emergence: Complexity and Organization,* 11 (4), pp. 15-25.

Lichtenstein, B., 2011. Complexity Science Contributions to the Field of Entrepreneurship. In: Allen, P., Maguire, S., McKelvey, B., eds., 2011. *The SAGE Handbook of Complexity and Management.* London: Sage, pp. 471-493.

Liguori, E., Winkler, C., Hechavarria, D., and Lange, J. 2018. Guest Editorial: Interdisciplinary perspectives on entrepreneurial ecosystems. *Journal of Enterprising Communities* 12.2: pp. 86-91.

Limpert, E., Stahel, W. A., and Abbt, M., 2001. Log-normal distributions across the sciences: Keys and clues. *Bioscience*, *51*, pp. 341–352.

Limpert E, Stahel, W.A., 2011. Problems with Using the Normal Distribution – and Ways to Improve Quality and Efficiency of Data Analysis. *PLoS ONE* 6, p. 7.

Lotti, F., Santarelli, E., and Vivarelli, M., 2009. Defending Gibrat's law as a long-run regularity. *Small Business Economics*, 32(1), pp. 31–44.

Lucas Jr, R. E., 1978. On the size distribution of business firms. *The Bell Journal of Economics*, Vol. 9, No. 2 (Autumn), pp. 508-523.

Lundström, A., Vikström, P., Fink, M., Meuleman, M., Głodek, P., Storey, D. and Kroksgård, A., 2014. Measuring the Costs and Coverage of SME and Entrepreneurship Policy: A Pioneering Study. *Entrepreneurship: Theory and Practice*, 38, pp. 941–957.

Luttmer, E. G., 2007. Selection, growth, and the size distribution of firms. *The Quarterly Journal of Economics,* 122(3), pp.1103-1144.

Macal, C.M., and North, M.J., 2009. Agent-based modeling and simulation. In: M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin and R. G. Ingalls, eds., *Proceedings of the 2009 Winter Simulation Conference (WSC)*, Austin, TX, USA, December 13-16, 2009, IEEE, pp.86-98. Available at:

http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5429318&isnumber=5429163

Macal, C.M., and North, M.J., 2010. Tutorial on agent-based modelling and simulation. *Journal of Simulation* 4, 151–162. Available at:

http://www2.econ.iastate.edu/tesfatsi/ABMTutorial.MacalNorth.JOS2010.pdf

Macal, C. M., 2016. Everything you need to know about agent-based modelling and simulation, *Journal of Simulation*, 10:2, 144-156.

Macy, M. W. and Willer, R., 2002. From Factors to Actors: Computational Sociology and Agent-Based Modelling. *Annual Review of Sociology*, 28, pp. 143–66.

Maguire, S., McKelvey, B., Mirabeau, L., and Otzas, N., 2006. Complexity Science and Organization Studies. In: Cleggs, S.R., et al., eds., *The SAGE Handbook of Organization Studies.* London: Sage, pp. 165-214.

Maguire, S. Allen, P., and McKelvey, B., 2011. Complexity and Management: Introducing the SAGE Handbook, in Allen, P., Maguire, S., McKelvey, B., eds., 2011. *The SAGE Handbook of Complexity and Management.* London: Sage. pp 1-26.

Malevergne, Y., Pisarenko, V., and Sornette, D., 2005. Empirical distributions of stock returns: Between the stretched exponential and the power law? *Quantitative Finance*, 5, pp. 379–401.

Malevergne Y, Pisarenko V, Sornette D., 2011. Gibrat's law for cities: uniformly most powerful unbiased test of the Pareto against the lognormal. *Physical Review* E, 83, 036111.

Mandelbrot, B., 1960. The Pareto-Lévy Law and the Distribution of Income, *International Economic Review*, vol. 1, no. 2, pp. 79-106.

Mandelbrot, B. B., 1963. The variation of certain speculative prices. *Journal of Business*, 36, pp. 394–419.

Mandelbrot, B. B., and Taleb, N. N. 2010. Mild vs. wild randomness: Focusing on those risks that matter. In: F. X. Diebold, N. A. Doherty, & R. J. Herring, Eds., *The known, the unknown and the unknowable in financial institutions* Princeton, NJ: Princeton University Press, pp. 47–58.

Mankiw, N. G., 2013. Defending the one percent. *Journal of Economic Perspectives*, 27, pp. 21–34.

Marsili, O., 2005. Technology and the Size Distribution of Firms: Evidence from Dutch Manufacturing. *Review Of Industrial Organization,* 27, no. 4 (December 2005), pp. 303-328.

Marsili, O., 2006. Stability and Turbulence in the Size Distribution of Firms: Evidence from Dutch Manufacturing. *International Review Of Applied Economics,* 20, no. 2, pp. 255-272.

MASON (2016). http://cs.gmu.edu/~eclab/projects/mason/

Martinez, M.A., Yang, T. and Aldrich, H.E., 2011. Entrepreneurship as an evolutionary process: research progress and challenges. *Entrepreneurship Research Journal*, Vol. 1, No. 1, pp. 1-28.

McAlister D., 1879. The law of the geometric mean. *Proceedings of the Royal Society*, 29, pp. 367–376.

McDonald, S., Gan, B.C., Fraser, S.S., Oke, A. and Anderson, A.R., 2015. A review of research methods in entrepreneurship 1985-2013. *International*

Journal of Entrepreneurial Behavior & Research, Vol. 21 No. 3, pp. 291-315.

McKelvey, B., 1999. Complexity Theory in Organization Science: Seizing the Promise or Becoming a Fad. *Emergence*, 1, 1, p. 5-32.

McKelvey, B., 2004. Toward a complexity science of entrepreneurship. *Journal of Business Venturing*, 19, pp. 313-341.

McKelvey, B., ed., 2011. *Complexity*. New York: Routledge.

McKelvey, B., Lichtenstein, B.B. and Andriani, P., 2012. When organisations and ecosystems interact: toward a law of requisite fractality in firms. *Int. J. Complexity in Leadership and Management,* Vol. 2, Nos. 1/2, pp.104–136.

McKelvie, A. and Wiklund, J., 2010. Advancing firm growth research: A focus on growth mode instead of growth rate. *Entrepreneurship: Theory and Practice,* 34, pp. 261–288.

McLane, A. J., Semeniuk, C., McDermid, G. J., and Marceau, D. J. 2011. The role of agent-based models in wildlife ecology and management. *Ecological Modelling*, *222*(8), 1544-1556.

McMullen, J. S. and Shepherd, D. A., 2006. Entrepreneurial action and the role of uncertainty in the theory of the entrepreneur. *Academy of Management Review,* 31,132–152.

McMullen, J.S. and Dimov, D., 2013. Time and the entrepreneurial journey: The problems and promise of studying entrepreneurship as a process. *Journal of Management Studies*, 50(8), pp. 1481-1512.

Mantegna, R. N. and Stanley, H. E., 1999. *An Introduction to Econophysics: Correlations and Complexity in Finance.* Cambridge: Cambridge University Press.

Merton, R. K., 1968. The Matthew effect in science. *Science,* 159, pp. 56–63.

Miller, G. A., 1957. Some effects of intermittent silence. *American Journal of Psychology* 70, 311–314.

Miller, J. H. and Page, S. E., 2007. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life.* Princeton, NJ: Princeton University Press.

Minniti, M., 2004. Entrepreneurial alertness and asymmetric information in a spin-glass model. *Journal of Business Venturing*, Volume 19, Issue 5, September 2004, pp. 637–658.

Mitchell, M., 2009. *Complexity: A Guided Tour*. Oxford: Oxford University Press.

Mitzenmacher, M., 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1, pp. 226–251.

Mitzenmacher, M., 2005. Editorial: The Future of Power Law Research. *Internet Mathematics*, 2:4, pp. 525-534.

Mossa, S., Barthélémy, M., Stanley, H. E., and Nunes Amaral, L. A., 2002. Truncation of power law behavior in “scale-free” network models due to information filtering. *Physical Review Letters*, 88, Article 138701.

Naudé, W., ed., 2010. *Entrepreneurship and Economic Development.* United Nations University World Institute for Development Economics Research (UNU-WIDER) Basingstoke: Palgrave Macmillan.

Nettle D., Gibson, M.A., Lawson D.W., Sear, R. 2013. Human behavioral ecology: current research and future prospects. *Behavioral Ecology*, 24(5), 1031–1040.

Newman, M. E. J. 2005. Power laws, Pareto distributions and Zipf’s law. *Contemporary Physics*, 46, pp. 323–351. Available at:

Power laws, Pareto distributions and Zipf's law.

Newman, M. E. J., 2011. Complex systems. *American Journal of Physics*, Vol 79, pp. 800-810.

Nicolaou N., and Shane S., 2009. Can genetic factors influence the likelihood of engaging in entrepreneurial activity? *Journal of Business Venturing*, 24, no. 1, pp. 1-22.

Nicolaou, N. and Shane, S., 2014, Biology, Neuroscience, and Entrepreneurship. *Journal of Management Inquiry,* vol. 23, no. 1, pp. 98-100.

Nicolis, G. and Rouvas-Nicolis, C., 2007. Complex systems. Scholarpedia, 2(11):1473. http://www.scholarpedia.org/article/Complex_systems

Nightingale, P. and Coad, A., 2014. Muppets and gazelles: political and methodological biases in entrepreneurship research. *Industrial and Corporate Change*, vol. 23, no. 1, pp. 113-143.

Nirei, M., and Souma, W., 2007. A two factor model of income distribution dynamics. *Review of Income and Wealth*, 53, pp. 440–459.

Nofal, A.M., Nicolaou, N., Symeonidou, N. and Shane, S., 2018, Biology and Management: A Review, Critique, and Research Agenda. *Journal of Management,* vol. 44, no. 1, pp. 7-31.

North, M.J., and Macal, C.M., 2007. *Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modelling and Simulation*. Oxford: Oxford University Press.

O'Boyle, E., Jr., and Aguinis, H., 2012. The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology*, 65, pp. 79–119.

Okuyama K., Takayasu M. and Takayasu H., 1999. Zipf's law in income distribution of companies. *Physica A,* 269:125–31.

OECD, 2007. *OECD framework for the evaluation of SME and entrepreneurship policies and programmes.* Paris: OECD Publishing.

Page, S.E., 2008. Agent-based models. In: Steven N. Durlauf and Lawrence E. Blume, Eds., *The New Palgrave Dictionary of Economics*. Second Edition. Basingstoke: Palgrave Macmillan.

Pareto V., 1896. *Cours d'Economie Politique*. Geneva: Droz.

Penrose, E., 1959. *Theory of the growth of the firm.* New York: Oxford University Press.

Pons Rotger, G., Gørtz, M., and Storey, D.J., 2012. Assessing the effectiveness of guided preparation for new venture creation and performance: Theory and practice. *Journal of Business Venturing*, 27(4), 506–521.

Price, D. J. de S., 1976. A general theory of bibliometric and other cumulative advantage processes. *J. Amer. Soc. Inform. Sci.* **27** (5), pp. 292–306.

Quandt, R. E. 1966. On the Size Distribution of Firms. *The American Economic Review*, Vol. 56, No. 3 (June), pp. 416-432.

R Core Team (2018) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. Available at:

https://www.r-project.org.

Raichlen, D.A., Wood, B.M., Gordon, A.D., Audax Z. P. Mabulla, Marlowe, F.W. and Pontzer, H. 2014. Evidence of Lévy walk foraging patterns in human hunter—gatherer", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 111, no. 2, pp. 728-733.

Railsback, S.F., and Grimm, V., 2012. *Agent-Based and Individual-Based Modelling: A Practical Introduction*. Princeton and Oxford: Princeton University Press [Second edition in early 2019].

Railsback, S. F., Ayllón, D., Berger, U., Grimm, V., Lytinen, S. L., Sheppard, C.J.R. and Thiele, J.C., 2017. Improving execution speed of models implemented in NetLogo. *Journal of Artificial Societies and Social Simulation.* **20** (1) 3. Available at:

http://jasss.soc.surrey.ac.uk/20/1/3.html

Rapaport, L.G. and Brown, G.R. 2008. Social influences on foraging behavior in young nonhuman primates: learning what, where and how to

eat. *Evolutionary Anthropology: Issues, News, and Reviews*. **17** (4): 189–201.

Repast Simphony 2.5 (*2017*) https://repast.github.io/

Reynolds, P., and White, S. 1997. *The Entrepreneurial Process.* Quorum Books: Westport, CT.

Reynolds, P., Bosma, N., Autio, E., Hunt, S. De Bono, N., Servais, I., Lopez-Garcia, P., and Chin, N., 2005. Global Entrepreneurship Monitor: Data Collection Design and Implementation 1998-2003. *Small Business Economics,* 24, no. 3, pp. 205-231.

Reynolds, P.D. and Curtin, R.T., 2008. Business Creation in the United States: Panel Study of Entrepreneurial Dynamics II Initial Assessment. *Foundations and Trends in Entrepreneurship,* 4, 3 (Jan. 2008), 155-307.

Reynolds, P. D., and Curtin, R.T.,Eds., 2010. *New Business Creation: An International Overview*. Series: International Studies in Entrepreneurship, Vol. 27. New York: Springer.

Reynolds, P. D., and Curtin, R. T., 2011. PSED I, II harmonized transitions, outcomes data set. Retrieved from: **http://www.psed.isr.umich.edu**

Reynolds P., Hart, M., and Mickiewicz, T., 2014. *The UK Business Creation Process: The 2013 Panel Study of Entrepreneurial Dynamics Pretest.* Birmingham, U.K.: Aston Business School, Enterprise Research Centre.

Reynolds, P. D., Hechavarria, D., Tain, L., Samuelsson, M., and Davidsson, P., 2016. Panel Study of Entrepreneurial Dynamics: A five cohort outcomes data set. Research Gate. Available at:

https://www.researchgate.net/publication/294292920_Panel_Study_of_Entrepreneurial_Dynamics_A_Five_Cohort_Outcomes_Harmonized_Data_Set?channel=doi&linkId=56bfe94908aeedba0562fc15&showFulltext=true

[Accessed 25 August 2018]

Reynolds, P. D., 2017a. When is a Firm Born?: Alternative Criteria and Consequences. *Business Economics*, 52(1), pp. 41-56.

Reynolds, P. D., 2017b. Tracking the Entrepreneurial Process with the Panel Study of Entrepreneurial Dynamics (PSED) Protocol. In: Oxford Research Encyclopedia, Business and Management. Oxford University Press, USA.

http://business.oxfordre.com/view/10.1093/acrefore/9780190224851.001.0001/acrefore9780190224851-e-86 [Accesed 26 March 2018]

Richerson P.J., Boyd R. and Bettinger R.L. 2001. Was agriculture impossible during the pleistocene but mandatory during the holocene? A climate change hypothesis. *American Antiquity,* 66:387‑411.

Ridley, M., 2004. *Evolution.* 3rd ed. Malden, MA: Blackwell.

.

Robertson, D.A. and Caldart, A.A., 2008. Natural science models in management: opportunities and challenges. *Emergence: Complexity and Organization*, Vol. 10 No. 2, pp. 61-75.

Rosser, J. Barkley, Jr., 2008. Econophysics. In: Steven N. Durlauf and Lawrence E. Blume, Eds., *The New Palgrave Dictionary of Economics*, Second Edition. Basingstoke: Palgrave Macmillan.

Roundy, P.T., Bradshaw, M. & Brockman, B.K., 2018. The emergence of entrepreneurial ecosystems: A complex adaptive systems approach. *Journal of Business Research,* vol. 86, pp. 1-10.

Rundle, J. B., Turcotte, D. L., Shcherbakov, R., Klein, W., and Sammis, C., 2003. Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. *Rev. Geophys*., 41, p. 1019.

Samuelsson. M., 2011. The Swedish PSED: performance in the nascent venturing process and beyond. In: *New business Creation: An International Overview*, Reynolds, P. and Curtin, R., eds., New York: Springer, pp. 223-253.

Sarasvathy, S.D., 2003. Entrepreneurship as a science of the artificial. *Journal of Economic Psychology*, Vol. 24 No. 2, pp. 203-20.

Sawyer, R.K., 2003. Artificial societies: multiagent systems and the micro-macro link in sociological theory. *Sociological Methods & Research*, Vol. 31 No. 3, pp. 325-63.

Schelling, T.C., 1971. Dynamic models of segregation. *Journal of Mathematical Sociology*, Vol. 1 No. 2, pp. 143-86.

Schelling, T., 2006. *Micromotives and macrobehavior* (New ed. with a new preface and the Nobel lecture. ed., Fels lectures on public policy analysis). New York ; London: W.W. Norton. [Previous ed.: 1978].

Schmolke A., Thorbek P., DeAngelis D.L., Grimm V., 2010. Ecological modelling supporting environmental decision making: a strategy for the future. *Trends in Ecology and Evolution* 25, pp. 479-486.

Schumpeter, J.A., 1934. *The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle.* Cambridge: Harvard University Press.

Schumpeter, J., 1949. Vilfredo Pareto (1848–1923). *The Quarterly Journal of Economics*, 63, pp. 147–73.

Shane, S. and Venkataraman, S., 2000. The promise of entrepreneurship as a field of research. *Academy of Management Review*, 25(1), pp. 217-226.

Shane, S. 2004. An evolving field: Guest editor's introduction to the special issue on Evolutionary Approaches to Entrepreneurship in honor of Howard Aldrich, *Journal of Business Venturing*, vol. 19, no. 3, pp. 309-312.

Shane S., 2009. Introduction to the focused issue on the biological basis of business. *Organizational Behavior and Human Decision Processes*, 110: 67-69.

Shane, S., 2012. Reflections on the 2010 AMR Decade Award: delivering on the promise of entrepreneurship as a field of research. *Academy of Management Review*, 37, pp. 10–20.

Shim, J., 2016. Toward a more nuanced understanding of long-tail distributions and their generative process in entrepreneurship. *Journal of Business Venturing Insights*, Vol 6, December, pp. 21-27.

Shim, J. Bliemel, M., and Choi, M., 2017. Modeling complex entrepreneurial processes: A bibliometric method for designing agent-based simulation models, *International Journal of Entrepreneurial Behavior & Research*, Vol. 23, Issue: 6, pp.1052-1070.

Shim, J. and Bliemel, M., 2018, Ignition of New Product Diffusion in Entrepreneurship: An Agent-Based Approach. *Entrepreneurship Research Journal*, vol. 8, no. 2, pp. 216-232.

Shim, J., and Davidsson, P., 2018. Shorter than we thought: The duration of venture creation processes. *Journal of Business Venturing Insights*, vol. 9, pp. 10-16.

Shockley, W., 1957. On the Statistics of Individual Variation of Productivity in Research Laboratories, *Proceedings of IRE* 45, Volume 45, Issue: 3, (March), pp. 279-90.

Simon, H. A., 1955a. On a class of skew distribution functions. *Biometrika* **42** (3–4), pp. 425–440.

Simon, H.A., 1955b. A behavioral model of rational choice. *Quarterly Journal of Economics*, Vol. 69 No. 1, pp. 99-118.

Simon, H.A., 1956. Rational choice and the structure of the environment. *Psychological Review*, Vol. 63 No. 2, pp. 129-38.

Simon, H. A., and Bonini, C. P., 1958. The size distribution of business firms. *The American Economic Review*, Vol. 48, No. 4 (Sep), pp. 607-617.

Simon, H. A., 1962. The architecture of complexity. *Proceedings of the American Philosophical Society,* Vol. 106, No. 6 (Dec. 12, 1962), pp. 467-482.

Simon, H.A., 1991. Bounded rationality and organizational learning. *Organization Science*, Vol. 2 No. 1, pp. 125-34.

Simon, H. A., 1996. *The Sciences of the Artificial.* Third edition. Cambridge, Massachusetts: MIT Press.

Singer, S., Amorós, J.E., and Moska, D., 2015. *Global Entrepreneurship Monitor. 2014 Global Report.* Global Entrepreneurship Research Association (GERA), London: London Business School.

http://www.gemconsortium.org/report

Sommer, S., Loch, C., and Dong, J., 2009. Managing complexity and unforeseeable uncertainty in start-up companies: An empirical study. *Organization Science*, 20, pp. 108-113.

Sorenson, O. and Stuart, T., 2008. Entrepreneurship: a field of dreams? *Academy of Management Annals*, 2, pp. 517–543.

Sornette, A. and Sornette, D., 1989. Self-organized criticality and earthquakes. *Europhys. Lett.* 9, pp. 197–202.

Sornette, D., 2004. *Why Stock Markets Crash: Critical Events in Complex Financial Systems.* Princeton: Princeton University Press.

Sornette, D., 2006. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Self-oganization and Disorder: Concepts and Tools*, 2nd ed., Berlin: Springer.

Sornette, D., and Ouillon, G., 2012. Dragon-kings: Mechanisms, statistical methods and empirical evidence. *European Physical Journal Special Topics*, 205, pp. 1–26.

Sporns, O., 2007. Complexity. *Scholarpedia*, 2(10):1623.

http://www.scholarpedia.org/article/Complexity

Stanley, M. H. R., Buldyrev, S. V., Havlin, S., Mantegna, R. N., Salinger, M. A., and Stanley, E. H., 1995. Zipf plots and the size distribution of firms, Economics Letters, 49, pp. 453–457.

Stanley, M.H.R., Amaral, L.A.N., Buldyrev, S.V., Havlin, S., Leschhorn, H., et al. 1996. Scaling behavior in the growth of companies. Nature, 379, pp. 804–6.

Stanley, H. E., Amaral, L., Gopikrishnan, P. and Plerou, V., 2000. Scale invariance and universality of economic fluctuations. *Physica A,* vol. 283(1-2) (August), pp. 31-41.

Stanley, H. E. and Plerou, V., 2001. Scaling and universality in economics: empirical results and theoretical interpretation. *Quantitative Finance*, vol. 1 (6) (November), pp. 563-7.

Stevenson, H. and Harmeling, S., 1990. Entrepreneurial management's need for a more “chaotic” theory. *Journal of Business Venturing*, Volume 5, Issue 1, pp. 1–14.

Stephens, D.W., and Krebs J.R. 1986. *Foraging theory*. Princeton, NJ: Princeton Univ. Press.

Stephens, D.W., Brown, J.S. and Ydenberg, R.C. 2007. *Foraging: Behavior and ecology*. Chicago: Univ. of Chicago Press.

Strogatz, S.H., 1994. *Nonlinear Dynamic and Chaos: with Applications to Physics, Biology, Chemistry and Engineering*. Reading (Massachusetts): Perseus Books.

Strogatz, S., 2004. *Sync: The Emerging Science of Spontaneous Order.* New York: Penguin.

Stumpf, M. P. H., and Porter, M. A., 2012. Critical truths about power laws. *Science*, 335 (6069), pp. 665–666.

J. Sutton, J., 2002. The variance of firm growth rates: the ‘Scaling’ puzzle. *Physica A*, 312, pp. 577–590.

Tao, Y., 2015. Universal laws of human society’s income distribution. *Physica A: Statistical Mechanics and its Applications*, vol. 435, pp. 89-94.

Tao, Y., Wu, X., Zhou, T., Yan, W., Huang, Y., Yu, H., Mondal, B. and Yakovenko, V.M., 2017. Exponential structure of income inequality: evidence from 67 countries. *Journal of Economic Interaction and Coordination*, December 2017, pp. 1-32.

Taleb, N. N., 2007. *The black swan: The impact of the highly improbable*. New York, NY: Random House.

Taleb, N. N., 2012. *Antifragile: Things that gain from disorder*. New York, NY: Random House.

Tasaka, D., 1999. Twenty-first-century Management and the Complexity Paradigm, *Emergence*, 1, 4, p. 115-123.

ten Broeke G, van Voorn G. and Ligtenberg A., 2016. Which sensitivity analysis method should I use for my agent-based model? *Journal of Artificial Societies and Social Simulation*, 19(1), p. 5. Available at:

http://jasss.soc.surrey.ac.uk/19/1/5.html

Tesfatsion, L., 2002. Agent based computational economics: growing economies from the bottom up. *Artificial Life*, Vol. 8 No. 1, pp. 55-82.

Tesfatsion, L., 2018. Agent-based Computational Economics (ACE) Growing Economies from the Bottom Up: Available at:

http://www2.econ.iastate.edu/tesfatsi/

[Accessed 15 August 2018]

Thiele, J.C. and Grimm, V., 2010. NetLogo meets R: Linking agent-based models with a toolbox for their analysis. *Environmental Modelling and Software,* Volume 25, Issue 8, pp. 972 – 974.

Thiele, J., Kurth, W., and Grimm, V., 2012a. RNetLogo: An R package for running and exploring individual-based models implemented in NetLogo. *Methods in Ecology and Evolution*, 3(3), pp. 480–483.

Thiele, J.C., Kurth, W. and Grimm, V., 2012b. Agent-Based Modelling: Tools for Linking NetLogo and r. *Journal of Artificial Societies and Social Simulation,* vol. 15, no. 3.

Thiele, J.C., 2014. R Marries NetLogo: Introduction to the RNetLogo Package. *Journal of Statistical Software,* vol. 58, no. 1, pp. 1-41.

Thiele, J.C., Kurth, W. and Grimm, V., 2014. Facilitating Parameter Estimation and Sensitivity Analysis of Agent-Based Models: A Cookbook Using NetLogo and 'R'. *Journal of Artificial Societies and Social Simulation,* vol. 17, no. 3, p. 11. Also available at:

http://jasss.soc.surrey.ac.uk/17/3/11.html

Thomas, L.W., and Autio, E., 2014 The Fifth Facet: The Ecosystem as an Organizational Field. *Academy of Management Annual Meeting Proceedings*, vol. 2014, no. 1, p. 1.

Tinbergen N. 1963. On aims and methods of ethology. *Zeitschrift*

fur Tierpsychologie, 20: 410–433.

Tucker, B. 2007. Applying behavioral ecology and behavioural economics to conservation and development planning: example from the Mikea Forest, Madagascar. *Human Nature*, 18: 190–208.

University of Michigan 2018. Panel Study of Entrepreneurial Dynamics.

Available at:

http://www.psed.isr.umich.edu

[Accessed 17 August 2018]

Uzzi, B., Amaral, L.A.N. and Reed-Tsochas, F., 2007. Small world networks and management science research: a review. *European Management Review*, Vol. 4 No. 2, pp. 77-91.

Van de Ven, A.H. and Engleman, R.M., 2004. Event and outcome-driven explanations of entrepreneurship. *Journal of Business Venturing*, Vol. 19 No. 3, pp. 343-58.

Virkar, Y. and Clauset, A., 2014. Power-law distributions in binned empirical data. *Annals of Applied Statistics* 8, no. 1, 89--119.

Vicsek, T., 2002. Complexity: The Bigger Picture. *Nature*, Vol. 418, p. 131 ff.

Villasenor-Alva, J.A. and Gonzalez-Estrada, E., 2009. A bootstrap goodness of fit test for the generalized Pareto distribution. *Computational Statistics and Data Analysis*, 53, 11, 3835-3841.

Villasenor, J.A. and Gonzalez-Estrada, E., 2015. A variance ratio test of fit for Gamma distributions. *Statistics and Probability Letters*, 96 1, 281-286.

Watts, D.J. and Strogats, S.H., 1998. Collective dynamics of small world networks. *Nature*, Vol. 393 No. 6684, pp. 440-2.

Watts, D.J., 1999. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton, NJ: Princeton University Press.

Weber, B. H., 2011. Extending and Expanding the Darwinian Synthesis: The Role of Complex Systems Dynamics. *Studies in History and Philosophy of Biological and Biomedical Science*. 42 (1): 75–81.

West, B. J. and Deering, B., 1995. *The Lure of Modern Science: Fractal Thinking.* Singapure: World Scientific Publishing Co.

West, G. B., James H. Brown, and Brian J. E., 1997. A general model for the origin of allometric scaling laws in biology. *Science*, 276 (5309), pp. 122—126.

West, G. B., 2017. *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies.* London: Weidenfeld & Nicolson.

Westhead, P. and Wright, M., 2013. *Entrepreneurship: A Very Short Introduction*. Oxford: Oxford University Press.

Whetten, D. A., Felin, T., and King, B. G., 2009. The practice of theory borrowing in organizational studies: Current issues and future directions. *Journal of Management*, *35*, pp. 537–563.

Wilensky, U., (1997). NetLogo Ants model.

http://ccl.northwestern.edu/netlogo/models/Ants

Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Wilensky, U. (1997). NetLogo Wolf Sheep Predation model.

http://ccl.northwestern.edu/netlogo/models/WolfSheepPredation

Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Wilensky, U. (1999) NetLogo.

http://ccl.northwestern.edu/netlogo/.

Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL.

Wilensky, U. and Shargel, B., 2002. BehaviorSpace. Evanston, IL: Center for Connected Learning and Computer Based Modeling, Northwestern University.

Wilensky, U. and Rand, R., 2015. *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo.* Cambridge, Massachusetts: MIT Press.

Wiklund, J., Davidsson, P., Audretsch, D. B. and Karlsson, C., 2011. The future of entrepreneurship research. *Entrepreneurship Theory and Practice,* 35, 1–9.

Willis, J.C. and Yule, G. U., 1922. Some statistics of evolution and geographical distribution in plants and animals, and their significance. *Nature,* 109, pp. 177–179.

Wolfram, S., 2002. *A New Kind of Science*. Champaign, IL: Wolfram Media.

Winterhalder B. P. 1981. Foraging strategies in the boreal forest: an analysis of Cree hunting and gathering In: Winterhalder B and Smith EA (eds) *Hunter-Gatherer Foraging Strategies,* pp. 66‑98. Chicago: University of Chicago Press.

Yang, S. S., and Chandra, Y., 2013. Growing artificial entrepreneurs: Advancing entrepreneurship research using agent-based simulation approach. *International Journal of Entrepreneurial Behaviour & Research*, Vol. 19 Iss: 2, pp. 210 – 237.

Yule, G. U., 1925. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. *Philos. Trans. R. Soc. London B* 213, pp. 21–87.

Zambrano, E., Hernando, A., Fernández Bariviera, A., Hernando, R., and Plastino, A. 2015. Thermodynamics of firms' growth. *Journal of the Royal Society Interface*, *12*(112), 20150789.

APPENDICES:

APPENDIX 1: TABLE 1 - “DISTRIBUTION PITTING STATISTICS - DPIT() RESULTS”

Table 1: Distribution Pitting Statistics (Dpit() Results)

Exploring log-normal distributions in nascent entrepreneurship outcomes: International comparisons and agent-based modelling.

by Ivan Rodriguez Hernandez (following pitting procedure and layout of Joo, Aguinis & Bradley, 2017)

Distribution Pitting Statistics Table for "Exploring log-normal distributions in nascent entrepreneurship outcomes: International comparisons and agent-based modelling."

The six columns of the table show the comparison results calculated by the software package Dpit() in R. For each comparison, it is shown the normalized log-likelihood ratio value followed by the normalized p-value (in parentheses). N is the sample.

Abbreviations of distribution names: PL = Pure power law, LogN = Lognormal, Exp = Exponential, Cut = Power law with an exponential cutoff, Norm = Normal, Pois = Poisson, and Weib = Weibull.

Abreviations of comparison between distributions: For example, NormvPL means Normal distribution versus power law distribution. A positive result of the normalized log-likelihood ratio value implies the first distribution indicates a superior fit in the comparison abbreviation name NormvPL. On the other hand, a negative result of the normalized log-likelihood ratio value implies that the second distribution is the superior fit.

p = statistical significance for the normalized log-likelihood ratio value.

Poisson's log-likelihood ratio and p-values are not available for continuous data.

Variable	*N*	NormvL	NormvCut PLvCut	NormvWeb PLvWeib CutvWeib	NormvLoN PLvLogN CutvLogN WiebvLoN	NormvExp PLvExp CutvExp WeibvExp LogNvExp	NormvPois PLvPois CutvPois WeibvPois LogNvPois ExpvPois
CAUSEE Australia - Outcome Variables *Full-Time Employees*							
1. Number of full-time Employees	**205**	**-4.96 (0)**	**-9.13 (0)**	**-9.01 (0)**	**-8.82 (0)**	**-11.18 (0)**	**3.44 (0)**
Young Firms – Wave 1 (Year 1)			**-40.54 (0)**	**-6.15 (0)**	**-5.78 (0)**	**-2.77 (0.006)**	**4.23 (0)**
Variable Name: W1: Q205#				**-1.26 (0.21)**	**-0.58 (0.56)**	**2.03 (0.04)**	**4.93 (0)**

					-0.19 (0.85)	**2.24 (0.025)**	**4.93 (0)**
						2.22 (0.03)	**4.94 (0)**
							5.04 (0)
2. Number of full-time Employees	**160**	**-6.19 (0)**	**-8.63 (0)**	**-8.60 (0)**	**-8.53 (0)**	**-11.51 (0)**	**2.94 (0.003)**
Young Firms – Wave 2 (Year 2)			**-20.40 (0)**	**-4.31 (0)**	**-4.11 (0)**	**0.05 (0.96)**	**3.76 (0.0002)**
Variable Name: W2_B16				**-1.02 (0.31)**	**-0.63 (0.53)**	**2.68 (0.007)**	**4.00 (0)**
					-0.12 (0.90)	**2.86 (0.004)**	**4.01 (0)**
						2.87 (0.004)	**4.01 (0)**
							4.04 (0)
3. Number of full-time Employees Young Firms – Wave 3 (Year 3) Variable Name: W3_B16	**127**	**-6.36 (0)**	**-8.21 (0)**	**-8.19 (0)**	**-8.22 (0)**	**--10.89 (0)**	**2.97 (0.003)**
			-14.46 (0)	**-3.85 (0.0001)**	**-3.62 (0)**	**1.04 (0.30)**	**3.50 (0)**
				-1.33 (0.18)	**-1.01 (0.31)**	**2.96 (0.003)**	**3.61 (0)**
					-0.45 (0.65)	**3.19 (0.001)**	**3.62 (0)**
						3.30 (0)	**3.62 (0)**
							3.61 (0)
4. Number of full-time Employees Young Firms – Wave 4 (Year 4) Variable Name: W4_B16	**100**	**-5.93 (0)**	**-6.73 (0)**	**-6.87 (0)**	**-7.01 (0)**	**-26.34 (0)**	**1.18 (0.24)**
			-6.35 (0)	**-2.83 (0.005)**	**-2.65 (0.008)**	**1.24 (0.21)**	**1.28 (0.20)**
				-4.26 (0)	**-3.13 (0.002)**	**1.53 (0.13)**	**1.28 (0.20)**
					-1.48 (0.14)	**1.70 (0.09)**	**1.28 (0.20)**
						1.77 (0.08)	**1.28 (0.20)**
							1.27 (0.20)
5. Number of full-time Employees **Young and Nascent Firms** – Wave 5 (Year 5) Variable Name: W5: Q24	**8,332**	**-18.06 (0)**	**-18.59 (0)**	**-18.64 (0)**	**-18.54 (0)**	**-19.95 (0)**	**-22.72 (0)**
			-46.04 (0)	**-3.15 (0)**	**-5.76 (0)**	**6.47 (0)**	**8.79 (0)**
				0.72 (0.47)	**-0.49 (0.62)**	**8.58 (0)**	**9.71 (0)**

					-1.68 (0.09)	**9.12 (0)**	**9.92 (0)**
						8.54 (0)	**9.67 (0)**
							10.23 (0)
Distribution Pitting Statistics (continued)							
6. Number of full-time Employees Nascent Firms – Wave 1 (Year 1) Variable Name: W1: Q252#	**73**	**-4.87 (0)**	**-5.59 (0)**	**-5.70 (0)**	**-5.82 (0)**	**-10.31 (0)**	**1.07 (0.29)**
			-3.86 (0)	**-1.88 (0.06)**	**-1.74 (0.08)**	**0.97 (0.33)**	**1.54 (0.12)**
				-2.19 (0.03)	**-1.65 (0.1)**	**1.42 (0.16)**	**1.57 (0.12)**
					-0.69 (0.49)	**1.59 (0.11)**	**1.58 (0.11)**
						1.68 (0.09)	**1.58 (0.11)**
							1.57 (0.11)
7. Number of full-time Employees Nascent Firms – Wave 2 (Year 2) Variable Name: W2_C79	**73**	**-4.08 (0)**	**-5.49 (0)**	**- -5.47 (0)**	**-5.73 (0)**	**-7.95 (0)**	**1.12 (0.26)**
			-8.36 (0)	**-2.96 (0)**	**-2.64 (0)**	**-0.06 (0.95)**	**1.71 (0.09)**
				-2.77 (0.17)	**-2.07 (0.04)**	**1.15 (0.25)**	**1.85 (0.06)**
					-1.34 (0.18)	**1.41 (0.16)**	**1.86 (0.06)**
						1.80 (0.07)	**1.89 (0.06)**
							1.88 (0.06)
8. Number of full-time Employees Nascent Firms – Wave 3 (Year 3) Variable Name: W3_C79	**52**	**-3.94 (0)**	**-6.08 (0)**	**-6.05 (0)**	**-6.30 (0)**	**--8.17 (0)**	**105.7 (0.08)**
			-8.16 (0)	**-3.38 (0)**	**-2.99 (0)**	**-0.52 (0.61)**	**2.23 (0.03)**
				-2.05 (0.04)	**-1.58 (0.11)**	**1.36 (0.17)**	**2.40 (0.02)**
					-1.17 (0.24)	**1.66 (0.10)**	**2.42 (0.02)**
						2.28 (0.02)	**2.45 (0.01)**
							2.44 (0.01)
9. Number of full-time Employees Nascent Firms – Wave 4 (Year 4) Variable Name: W4_C79	**48**	**-5.51 (0)**	**-7.08 (0)**	**-7.31 (0)**	**--7.93 (0)**	**-12.29 (0)**	**1.53 (0)**

			-4.70 (0.002)	**-2.74 (0.0062)**	**-2.27 (0.02)**	**0.76 (0.45)**	**1.83 (0.07)**
				-2.96 (0.004)	**-1.79 (0.07)**	**1.54 (0.12)**	**1.88 (0.06)**
					-1.02 (0.31)	**1.88 (0.06)**	**1.89 (0.06)**
						2.34 (0.02)	**1.91 (0.06)**
							1.89 (0.06)
10. Number of full-time Employees **Young and Nascent Firms** – Wave 5 (Year 5) Variable Name: W5_Q24 [same variable than YF]	**155**	**-4.52 (0)**	**-5.36 (0)**	**-5.25 (0)**	**-5.20 (0)**	**-6.57 (0)**	**- 2.12 (0.03)**
			-16.02 (0)	**-3.56 (0)**	**-3.53 (0)**	**1.13 (0.26)**	**2.44 (0.01)**
				-0.73 (0.47)	**-0.46 (0.65)**	**2.51 (0.01)**	**2.51 (0.01)**
					0.35 (0.72)	**2.49 (0.01)**	**2.50 (0.01)**
						2.42 (0.02)	**2.50 (0.01)**
							2.50 (0.01)
Total Sales							
11. Sales in $ (Total) (Last 12 Months) **Young Firms** – Wave 1 (Year 1) Variable Name: W1 Q2027#	**589**	**-9.29 (0)**	**-11.97 (0)**	**-9.66 (0)**	**-11.73 (0)**	**-25.41 (0)**	**1.92 (0.05)**
			-599.9 (0)	**-3.92 (0)**	**-22.31 (0)**	**2.08 (0.04)**	**1.92 (0.05)**
				54.16 (0.81)	**5.52 (0)**	**4.65 (0)**	**1.92 (0.05)**
					-19.13 (0)	**2.36 (0.02)**	**1.92 (0.05)**
						4.22 (0)	**1.92 (0.05)**
							1.91 (0.05)
12. Sales in $ (Total) (Last 12 Months) Young Firms – Wave 2 (Year 2) Variable Name: W2_B18	**483**	**-8.47 (0)**	**-14.08 (0)**	**-11.29 (0)**	**-13.18 (0)**	**-18.22 (0)**	**4.54 (0)**
			-492.7 (0)	**-10.98 (0)**	**-19.11 (0)**	**0.17 (0.86)**	**4.53 (0)**
				35.33 (0)	**6.54 (0)**	**7.72 (0)**	**4.53 (0)**
					-10.14 (0)	**3.47 (0)**	**4.54 (0)**
						6.58 (0)	**4.54 (0)**
							4.54 (0)
13. Sales in $ (Total) (Last 12 Months) Young Firms – Wave 3 (Year 3) Variable Name: W3_B18	**385**	**-8.64 (0)**	**-19.12 (0)**	**-14.78 (0)**	**-16.97 (0)**	**-22.91 (0)**	**5.84 (0)**
			-497.6 (0)	**-15.26 (0)**	**-22.49 (0)**	**-1.53 (0.12)**	**5.84 (0)**
				-25.72 (0)	**6.67 (0)**	**9.45 (0)**	**5.84 (0)**

					-8.36 (0)	**4.15 (0)**	**5.84 (0)**
						7.97 (0)	**5.84 (0)**
							5.84 (0)
14. Sales in $ (Total) (Last 12 Months) Young Firms – Wave 4 (Year 4) Variable Name: W4_B18	**298**	**-8.09 (0)**	**-14.67 (0)**	**11.14 (0)**	**13.10 (0)**	**-17.97 (0)**	**4.14 (0)**
			-362.88 (0)	**-9.18 (0)**	**-16.56 (0)**	**0.21 (0.83)**	**4.14 (0)**
				22.15 (0)	**6.64 (0)**	**7.62 (0)**	**4.14 (0)**
					-7.96 (0)	**3.11 (0.002)**	**4.14 (0)**
						6.02 (0)	**4.14 (0)**
							4.14 (0)
15. Sales in $ (Total) (Last 12 Months) Young Firms – Wave 5 (Year 5) Variable Name: W5_Q18 [&R32] [**note: same as NF**]	**393**	**-9.64 (0)**	**-13.76 (0)**	**-10.42 (0)**	**12.65 (0)**	**-20.32 (0)**	**2.79 (0.005)**
			-458.2 (0)	**- 4.00 (0)**	**-15.53 (0)**	**2.78 (0.005)**	**2.79 (0.005)**
				33.07 (0)	**7.67 (0)**	**7.13 (0)**	**2.78 (0.005)**
					-13.43 (0)	**3.40 (0)**	**2.79 (0.005)**
						5.91 (0.02)	**2.79 (0.005)**
							2.79 (0.005)
16. Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 1 (Year 1) Variable Name: W1 Q2030#	**302**	**-6.59 (0)**	**-8.94 (0)**	**-7.24 (0)**	**-8.57 (0)**	**-12.59 (0)**	**2.41 (0.016)**
			-138.8 (0)	**-3.70 (0.0002)**	**-14.27 (0)**	**1.82 (0.07)**	**2.41 (0.016)**
				6.08 (0)	**-3.61 (0.0003)**	**4.47 (0)**	**2.41 (0.016)**
					-11.09 (0)	**2.56 (0.01)**	**2.41 (0.016)**
						4.55 (0)	**2.41 (0.016)**
							2.41 (0.016)
17. Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 2 (Year 2) Variable Name: W2_C85_consolidated	**291**	**-7.06 (0)**	**-9.76 (0)**	**-8.84 (0)**	**-10.67 (0)**	**-20.35 (0)**	**2.61 (0.009)**

			-224.2 (0)	**-9.92 (0)**	**-17.15 (0)**	**0.67 (0.51)**	**2.61 (0.009)**
				31.4 (0)	**-1.94 (0.05)**	**3.73 (0.0002)**	**2.61 (0.009)**
					-10.04 (0)	**2.46 (0.014)**	**2.61 (0.009)**
						4.60 (0)	**2.61 (0.009)**
							2.61 (0.009)

Distribution Pitting Statistics (continued)

18. Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 3 (Year 3) Variable Name: W3_C85	**228**	**-7.52 (0)**	**-11.82 (0)**	**10.43 (0)**	**-12.18 (0)**	**-17.05 (0)**	**3.87 (0)**
			-181.7 (0)	**-7.55 (0)**	**-13.6 (0)**	**0.91 (0.37)**	**3.87 (0.0001)**
				30.38 (0)	**0.89 (0.38)**	**5.43 (0)**	**3.87 (0.0001)**
					-7.57 (0)	**3.47 (0.0005)**	**3.87 (0.0001)**
						6.07 (0)	**3.87 (0.0001)**
							3.87 (0.0001)
19. Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 4 (Year 4) Variable Name: W4_C85_consolidated	**159**	**-5.31 (0)**	**-9.42 (0)**	**-8.47 (0)**	**-9.26 (0)**	**-13.31 (0)**	**4.00 (0)**
			-113.9 (0)	**-5.23 (0)**	**-8.77 (0)**	**-0.56 (0.58)**	**3.99 (0)**
				14.40 (0)	**2.64 (0.008)**	**4.20 (0)**	**4.00 (0)**
					-4.04 (0)	**2.27 (0.02)**	**3.99 (0)**
						4.02 (0)	**3.99 (0)**
							3.99 (0)
20. Sales in $ (Total) (Last 12 Months) **Nascent and Young Firms – Wave 5 (Year 5)** Variable Name: W5_Q18[& R32] **Same variable than YF -15**	393	-12.98 (0)	-13.29 (0)	-13.57 (0)	-13.35 (0)	-15.39 (0)	-21.66 (0)
			-7.43 (0)	-0.56 (0.57)	-2.11 (0.04)	4.35 (0)	5.35 (0)
				0.75 (0.45)	-0.7 (0.48)	5.02 (0)	5.61 (0)
					-1.7 (0.09)	5.6 (0)	5.85 (0)
						5.17 (0)	5.67 (0)
							5.86 (0)
SWEDISH PSED - Outcome	**101**	**-6.66 (0)**	**-7.63 (0)**	**-7.78 (0)**	**-7.93 (0)**	**-10.99 (0)**	**2.13 (0.03)**

Variables *Full-Time Employees*							
21. Number of full-time Employees – SWE PSED 1 Wave 1 (Year 0) Variable Name: gw31nn00			**-7.05 (0.0002)**	**-3.17 (0.002)**	**-2.63 (0.009)**	**1.88 (0.06)**	**2.44 (0.01)**
				-2.47 (0.01)	**-1.82 (0.07)**	**2.63 (0.009)**	**2.47 (0.01)**
					-0.81 (0.41)	**2.93 (0.003)**	**2.48 (0.01)**
						3.16 (0.002)	**2.49 (0.01)**
							2.46 (0.01)
22. Number of full-time Employees– SWE PSED 1 Wave 2 (6 months) Variable Name: gw31nn06	**86**	**-5.25 (0)**	**-7.14 (0)**	**-7.20 (0)**	**-7.69 (0)**	**-11.34 (0)**	**1.47 (0.14)**
			-9.86 (0)	**-3.41 (0)**	**-2.96 (0.003)**	**0.10 (0.92)**	**2.07 (0.04)**
				-3.71 (0)	**-2.44 (0.01)**	**1.41 (0.15)**	**2.19 (0.03)**
					-1.54 (0.12)	**1.76 (0.08)**	**2.21 (0.03)**
						2.32 (0.02)	**2.24 (0.02)**
							2.23 (0.03)
23. Number of full-time Employees – SWE PSED 1 Wave 3 (12 months) Variable Name: gw31nn12	**61**	**-9.73 (0)**	**-9.89 (0)**	**-11.80 (0)**	**-10.75 (0)**	**-7.05 (0)**	**1.21 (0.23)**
			-0.35 (0.40)	**1.00 (0.32)**	**-0.65 (0.52)**	**2.41 (0.02)**	**1.25 (0.21)**
				4.96 (0)	**-0.58 (0.56)**	**2.44 (0.01)**	**1.25 (0.21)**
					-0.58 (0.56)	**1.46 (0.15)**	**1.24 (0.21)**
						2.58 (0.01)	**1.25 (0.21)**
							1.24 (0.21)

Distribution Pitting Statistics (continued)

24. Number of full-time Employees – SWE PSED 1 Wave 4 (18 months) Variable Name: gw31nn18	**57**	**-11.75 (0)**	**-11.95 (0)**	**-9.65 (0)**	**-13.38 (0)**	**-6.01 (0)**	**1.30 (0.19)**
			-0.44 (0.35)	**6.65 (0)**	**-0.76 (0.45)**	**3.24 (0.001)**	**1.31 (0.19)**
				- 6.85 (0.15)	**-0.66 (0.50)**	**3.27 (0.001)**	**1.31 (0.19)**
					-6.27 (0)	**1.94 (0.05)**	**1.31 (0.19)**
						3.54 (0)	**1.31 (0.19)**
							1.31 (0.19)
25. Number of full-time Employees – SWE PSED 1 Wave 5 (24 months) Variable Name: gw31nn24	**53**	**-6.50 (0)**	**-8.09 (0)**	**-8.43 (0)**	**-8.39 (0)**	**-12.63 (0)**	**2.11 (0.03)**
			-4.00 (0.005)	**-1.78 (0.08)**	**-1.75 (0.08)**	**1.57 (0.12)**	**2.50 (0.01)**
				-0.41 (0.68)	**-0.39 (0.70)**	**2.39 (0.02)**	**2.53 (0.01)**
					-0.39 (0.70)	**-0.26 (0.79)**	**2.64 (0.01)**
						2.54 (0.01)	**2.62 (0.01)**
							2.54 (0.01)
26. Number of full-time Employees – SWE PSED 1 Wave N75 (75 months) Variable Name: gw31n	**40**	**-10.45 (0)**	**-10.79 (0)**	**-9.31 (0)**	**-13.21 (0)**	**-5.73 (0)**	**1.27 (0.20)**
			-0.51 (0.31)	**4.46 (0)**	**-0.82 (0.41)**	**2.50 (0.01)**	**1.30 (0.19)**
				4.19 (0)	**-0.74 (0.46)**	**2.55 (0.01)**	**1.30 (0.19)**
					-4.17 (0)	**1.58 (0.11)**	**1.30 (0.19)**
						2.95 (0.003)	**1.31 (0.19)**
SWEDISH PSED - Outcome Variables							**1.30 (0.20)**
SALES TURNOVER (THOUSAND SEK)	**189**	**-7.43 (0)**	**-10.34 (0)**	**-7.76 (0)**	**-9.58 (0)**	**-17.1 (0)**	**1.52 (0.13)**
27. Sales Turnover (Thousands SEK) Last Year Variable Name: pt11nn18			**-81 (0)**	**-1.33 (0.18)**	**-10.66 (0)**	**1.73 (0.08)**	**1.52 (0.13)**
				4.02 (0)	**-3.58 (0.0003)**	**3.05 (0.002)**	**1.51 (0.13)**
					-11.26 (0)	**1.87 (0.06)**	**1.52 (0.13)**

						3.23 (0.001)	**1.52 (0.13)**
							1.52 (0.13)
	154	**-5.89 (0)**	**-10.7 (0)**	**-8.88 (0)**	**-10.06 (0)**	**-16.61 (0)**	**3.32 (0)**
28. Sales Turnover (Thousands SEK)							
First 3 Months							
Variable Name: pt12nn18			**-77.1 (0)**	**-5.55 (0)**	**-9.96 (0)**	**-0.37 (0.71)**	**3.33 (0)**
				2.42 (0.02)	**-2.97 (0)**	**3.51 (0)**	**3.33 (0)**
					-4.76 (0)	**2.00 (0.05)**	**3.33 (0)**
						3.84 (0)	**3.34 (0)**
							3.33 (0)
	151	**-4.67 (0)**	**-8.67 (0)**	**6.98 (0)**	**-7.91 (0)**	**-13.04 (0)**	**2.82 (0.005)**
29. Sales Turnover (Thousands SEK)							
First 6 Months			**-88.2 (0)**	**-6.29 (0)**	**-10.9 (0)**	**-0.65 (0.52)**	**2.83 (0)**
Variable Name: pt13nn18				**2.44 (0.01)**	**-2.13 (0.03)**	**3.31 (0)**	**2.83 (0)**
					-4.83 (0)	**1.78 (0.08)**	**2.83 (0)**
						3.30 (0)	**2.83 (0.005)**
							2.83 (0.005)

Distribution Pitting Statistics (continued)

30 Sales Turnover (Thousands SEK)	**13**	**-0.22 (0.83)**	**-3.08 (0)**	**-3.05 (0)**	**-2.58 (0.01)**	**-5.63 (0)**	**1.73 (0.08)**
First 12 Months			**-11.56 (0)**	**-2.08 (0.04)**	**-1.91 (0.06)**	**-1.36 (0.17)**	**1.73 (0.08)**
Variable Name: pt14nn18				**0.81 (0.42)**	**1.26 (0.21)**	**0.67 (0.50)**	**1.73 (0.08)**
					0.51 (0.61)	**0.22 (0.82)**	**1.73 (0.08)**
						-0.02 (0.98)	**1.73 (0.08)**
							1.73 (0.08)
31. Sales Turnover (Thousands SEK)							
Second year of operation (24 months)	**163**	**-4.35 (0)**	**-8.86 (0)**	**-7.58 (0)**	**-8.18 (0)**	**-11.99 (0)**	**3.56 (0)**
Variable Name: pt11nn24 (global dataset)							
			-91.2 (0)	**-7.27 (0)**	**-9.89 (0)**	**-1.73 (0.08)**	**3.57 (0)**
				1.69 (0.09)	**-2.07 (0.04)**	**3.16 (0)**	**3.57 (0)**

					-3.61 (0)	**1.85 (0.06)**	**3.57 (0)**
						3.35 (0)	**3.57 (0)**
							3.56 (0)
32. Sales Turnover (Thousands SEK) Sales Turnover in 1997 Variable Name: pt31nn24 (global dataset)	**14**	**-2.03 (0.04)**	**-3.10 (0)**	**-3.49 (0)**	**-2.85 (0)**	**-5.69 (0)**	**1.80 (0.07)**
			-2.65 (0.02)	**0.02 (0.99)**	**-0.93 (0.35)**	**0.46 (0.65)**	**1.80 (0.07)**
				1.16 (0.25)	**1.91 (0.06)**	**1.25 (0.21)**	**1.80 (0.07)**
					-0.71 (0.47)	**0.77 (0.44)**	**1.80 (0.07)**
						1.01 (0.31)	**1.80 (0.07)**
							1.80 (0.07)
33. Sales Turnover (Thousands SEK) Sales Turnover in 1998 Variable Name: pt21nn24 (global dataset)	**111**	**-3.79 (0)**	**-6.46 (0)**	**-5.63 (0)**	**-5.91 (0)**	**-8.29 (0)**	**3.11 (0.002)**
			-45.30 (0)	**-45.30 (0.003)**	**-5.68 (0)**	**-0.35 (0.73)**	**3.11 (0.002)**
				2.53 (0.01)	**0.18 (0.86)**	**2.91 (0.004)**	**3.11 (0.002)**
					-2.47 (0.01)	**1.61 (0.11)**	**3.11 (0.002)**
						2.54 (0.01)	**3.11 (0.002)**
							3.11 (0.002)
34. Sales Turnover (Thousands SEK) Last Year Sales Turnover after 75 months. Variable Name: pt11n (N75 SPSS file)	**123**	**-4.80 (0)**	**-8.29 (0)**	**-5.87 (0)**	**-7.13 (0)**	**-25.22 (0)**	**1.41 (0.15)**
			-71.4 (0)	**-5.34 (0.1)**	**-9.93 (0)**	**0.19 (0.86)**	**1.41 (0.16)**
				1.74 (0.08)	**-2.14 (0.03)**	**1.96 (0.05)**	**1.41 (0.16)**
					-6.61 (0)	**1.02 (0.31)**	**1.41 (0.16)**
						2.07 (0.04)	**1.41 (0.16)**
							1.41 (0.16)
35. Sales Turnover (Thousands SEK) Second year of operation (24 months) file SPSS erc-n24 Variable Name: SWE_pt11nn24_erc-n24	**171**	**-4.35 (0)**	**-8.83 (0)**	**-7.56 (0)**	**-8.15 (0)**	**-11.85 (0)**	**3.60 (0)**
			-95.7 (0)	**-7.47 (0)**	**-10.2 (0)**	**-1.80 (0.07)**	**3.60 (0)**
				1.70 (0.09)	**-2.09 (0.04)**	**3.21 (0)**	**3.60 (0)**
					-3.65 (0)	**1.87 (0.06)**	**3.60 (0)**
						3.38 (0)	**3.60 (0)**

							3.60 (0)
Distribution Pitting Statistics (continued)							
36. Sales Turnover (Thousands SEK) Sales Turnover in 1998 Variable Name: pt21nn24_erc-n24 – see file SPSS erc-n24	**113**	**-3.70 (0)**	**-6.40 (0)**	**-5.56 (0)**	**-5.84 (0)**	**-8.12 (0)**	**3.16 (0)**
SWE_pt21nn24_erc-n24			**-47.03 (0)**	**-3.18 (0)**	**-5.75 (0)**	**-0.46 (0.65)**	**3.16 (0)**
				2.52 (0.01)	**0.24 (0.81)**	**2.92 (0)**	**3.16 (0)**
					-2.40 (0.02)	**1.63 (0.10)**	**3.16 (0)**
						2.52 (0.01)	**3.16 (0)**
							3.16 (0)
37. Number of full-time Employees – SWE PSED 1 Wave 5 (24 months) Variable Name: gw31nn24 – Specific dataset erc-n24	**58**	**-7.65 (0)**	**-9.54 (0)**	**-10.0 (0)**	**-10.1 (0)**	**-12.5 (0)**	**- 2.49 (0.01)**
			-4.32 (0.003)	**-1.99 (0.05)**	**-1.89 (0.06)**	**1.94 (0.05)**	**2.91 (0.003)**
				-0.66 (0.51)	**-0.54 (0.59)**	**2.84 (0.005)**	**2.93 (0.003)**
					-0.07 (0.94)	**3.21 (0.001)**	**2.95 (0.003)**
						3.26 (0.001)	**2.95 (0.003)**
							2.91 (0.004)
38. Sales Turnover (Thousands SEK) Sales Turnover in 1997 Variable Name: pt31nn24_erc-n24 – see also file SPSS erc-n24	**16**	**-1.96 (0.05)**	**-3.29 (0)**	**-4.07 (0)**	**-3.01 (0)**	**-4.97 (0)**	**2.61 (0)**
			-3.64 (0)	**-0.20 (0.84)**	**-1.10 (0.27)**	**0.35 (0.73)**	**2.62 (0)**
				1.20 (0.23)	**2.03 (0.04)**	**1.41 (0.16)**	**2.62 (0)**
					-0.67 (0.50)	**0.96 (0.33)**	**2.62 (0)**
						1.14 (0.25)	**2.62 (0)**
							2.62 (0)
	119	**-8.47 (0)**	**-11.8 (0)**	**-11.37 (0)**	**-12.17 (0)**	**-16.5 (0)**	**- (0)**

39. PSED II USA Total Revenues BV2							
			-31.7 (0)	**-6.39 (0)**	**-6.57 (0)**	**1.63 (0.10)**	**- (0)**
				2.93 (0.5)	**-4.64 (0)**	**3.11 (0)**	**- (0)**
					-9.86 (0)	**2.92 (0)**	**- (0)**
						3.87 (0)	**- (0)**
							- (0)
	132	**-5.26 (0)**	**-6.76 (0)**	**-6.00 (0)**	**-7.18 (0)**	**-16.24 (0)**	**1.69 (0.09)**
40. PSED II USA Total Revenues CV2							
			-81.6 (0)	**-2.84 (0)**	**-8.03 (0)**	**0.77 (0.44)**	**1.69 (0.09)**
				20.2 (0)	**0.64 (0.52)**	**2.39 (0.02)**	**1.69 (0.09)**
					-6.29 (0.55)	**1.34 (0.18)**	**1.69 (0.09)**
						2.58 (0)	**1.69 (0.09)**
							1.69 (0.09)
	126	**-9.11 (0)**	**-10.6 (0)**	**-9.44 (0)**	**-12.0 (0)**	**-15.25 (0)**	**2.08 (0.04)**
41. PSED II USA Total Revenues DV2							
			-50.7 (0)	**1.31 (0.19)**	**-5.45 (0)**	**2.19 (0.03)**	**2.08 (0.04)**
				22.3 (0)	**0.52 (0.61)**	**3.21 (0)**	**2.08 (0.04)**
					-6.58 (0)	**1.97 (0.05)**	**2.08 (0.04)**
						3.65 (0)	**2.08 (0.04)**
							2.08 (0.04)

Distribution Pitting Statistics (continued)

42. PSED II USA Total Revenues EV2	**142**	**-8.61 (0)**	**-11.23 (0)**	**-9.92 (0)**	**-12.25 (0)**	**16.10 (0)**	**2.66 (0)**
			-82.3 (0)	**-1.73 (0.08)**	**-8.26 (0)**	**1.95 (0.05)**	**2.66 (0)**
				29.08 (0.83)	**0.70 (0.49)**	**3.96 (0)**	**2.66 (0)**
					-6.75 (0.57)	**2.44 (0.01)**	**2.66 (0)**
						4.56 (0)	**2.66 (0)**

							2.66 (0)
	135	**-6.49 (0)**	**-8.54 (0)**	**-8.01 (0)**	**-9.26 (0)**	**-17.22 (0)**	**- 2.75 (0)**
43. PSED II USA Total Revenues FV2							
			-60.6 (0)	**-1.95 (0.05)**	**-6.52 (0)**	**1.06 (0.29)**	**2.75 (0)**
				10.26 (0)	**0.60 (0.55)**	**3.07 (0)**	**2.75 (0)**
					-5.17 (0)	**1.91 (0.06)**	**2.75 (0)**
						3.50 (0)	**2.75 (0)**
							2.75 (0)
	44	**-1.53 (0)**	**-3.99 (0)**	**-3.99 (0)**	**-3.80 (0)**	**-4.43 (0)**	**-0.18 (0.86)**
44. PSED II USA Number regular Employees BU2							
Variable: BU2			**-8.76 (0)**	**-2.60 (0.009)**	**-2.44 (0.01)**	**-2.16 (0.03)**	**0.94 (0.35)**
				-0.004 (0.99)	**0.49 (0.62)**	**0.34 (0.74)**	**1.99 (0.05)**
					0.59 (0.55)	**0.37 (0.71)**	**1.99 (0.05)**
						-0.11 (0.91)	**1.96 (0.05)**
							2.07 (0.04)
	47	**-3.11 (0)**	**-4.26 (0)**	**-4.24 (0)**	**-4.11 (0)**	**-6.43 (0)**	**1.08 (0.28)**
45. PSED II USA Number regular Employees CU2							
			-6.20 (0)	**-2.25 (0.02)**	**-2.13 (0.03)**	**-0.03 (0.98)**	**1.40 (0.16)**
				-1.55 (0.12)	**-1.12 (0.26)**	**0.99 (0.32)**	**1.49 (0.14)**
					-0.61 (0.54)	**1.13 (0.26)**	**1.49 (0.14)**
						1.20 (0.23)	**1.50 (0.13)**
							1.50 (0.13)
	44	**-7.04 (0)**	**-8.16 (0)**	**-8.65 (0)**	**-9.09 (0)**	**-11.52 (0)**	**1.66 (0.097)**
46. PSED II USA Number regular Employees DU2							
			-2.47 (0.03)	**-2.10 (0.04)**	**-1.67 (0.09)**	**1.70 (0.089)**	**1.90 (0.06)**
				-2.00 (0.045)	**-1.19 (0.23)**	**2.12 (0.03)**	**1.91 (0.06)**
					-0.17 (0.86)	**2.45 (0.01)**	**1.92 (0.05)**
						2.61 (0)	**1.92 (0.05)**
							1.90 (0.06)

47. PSED II USA Number regular Employees EU2	**50**	**-9.87 (0)**	**-10.07 (0)**	**-8.62 (0)**	**-11.11 (0)**	**-6.83 (0)**	**1.27 (0.20)**
			-0.41 (0.36)	**4.67 (0)**	**-0.64 (0.52)**	**2.64 (0)**	**1.31 (0.19)**
				4.97 (0)	**-0.55 (0.59)**	**2.67 (0)**	**1.31 (0.19)**
					-4.65 (0)	**1.55 (0.12)**	**1.30 (0.19)**
						2.88 (0.004)	**1.31 (0.19)**
							1.30 (0.19)

Distribution Pitting Statistics (continued)

48. PSED II USA Number regular Employees FU2	**52**	**-9.60 (0)**	**-9.76 (0)**	**-8.72 (0)**	**-10.28 (0)**	**-7.06 (0)**	**1.24 (0.21)**
			-0.29 (0.45)	**4.43 (0)**	**-0.43 (0.67)**	**2.62 (0.009)**	**1.30 (0.20)**
				5.05 (0)	**-0.27 (0.78)**	**2.65 (0.008)**	**1.30 (0.20)**
					-5.17 (0)	**1.52 (0.13)**	**1.28 (0.20)**
						2.75 (0.006)	**1.30 (0.20)**
							1.28 (0.20)

APPENDIX 2: TABLE 2 - “DISTRIBUTION PITTING CONCLUSIONS”.

Table 2: Distribution Pitting Conclusions after Implementing the First, First Two, or All Three Decision Rules of Joo, Aguinis & Bradley (2017).

The columns of the table show the conclusions after Implementing the First, First Two, or All Three Decision Rules of Joo, Aguinis & Bradley (2017), using the results calculated by the software package Dpit() in R. For each comparison, it is shown the normalized *log-likelihood ratio value followed by the normalized p-value (in parentheses).*

Abbreviations of distribution names: PL = Pure power law, LogN = Lognormal, Exp = Exponential, Cut = Power law with an exponential cutoff, Norm = Normal, Pois = Poisson, and Weib = Weibull.

Abreviations of comparison between distributions: For example, ***NormvPL*** *means Normal distribution* ***versus*** *power law distribution. A positive result of the normalized log-likelihood ratio value implies the first distribution indicates a superior fit in the comparison abbreviation name* ***NormvPL****. On the other hand, a negative result of the normalized log-likelihood ratio value implies that the second distribution is the superior fit.*

p = statistical significance for the normalized log-likelihood ratio value. Hypothesis 0 is no statistical difference (p=1). The result is statistical significant with a low value (authors considered p<0.10; Joo, Aguinis & Bradley (2017).

Poisson's log-likelihood ratio and p-values are not available for continuous data.

ID	Variable	Distribution Pitting Decision After rule **1**	After decision **Rule 1 and 2**	After all three Decision rules **Rules 1, 2 & 3**	Comments on decisions
1	CAUSEE Australia Number of full-time Employees Young Firms – Wave 1 (Year 1) Variable Name: W1: Q205#	Undetermined Lognormal	Undetermined Lognormal	Cut (forcing inflexible instead lognormal)	CutvWeib: -1.258947 – p=0.2080495 CutvLogN -0.5771123 p = 0.5638636 WiebvLogN -0.1936912 p = 0.8464177 P values too high No rejection of three: Cut, Weibull, LogN Rule #3 forces to inflexible distribution: Cut
2	CAUSEE Australia Number of full-time Employees	Undetermined Lognormal	Undetermined	Cut (forcing inflexible	CutvWeib: -1.02 (0.31) CutvLogN: -0.63 (0.53) WiebvLogN : -0.12 (0.90)

	Young Firms – Wave 2 (Year 2) Variable Name: W2_B16			instead lognormal)	P value too high No rejection of three Cut, Weibull, LogN: rule #3 forces to inflexible distribution: Cut
3	CAUSEE Australia Number of full-time Employees Young Firms – Wave 3 (Year 3) Variable Name: W3_B16	Undetermined Lognormal (p very high)	Undetermined	Cut (forcing inflexible instead of log-normal)	CutvWeib: -1.33 (0.18) CutvLogN: -0.63 (0.53) WiebvLogN: -0.12 (0.90) No rejection of three: Cut, Weibull, LogN Rule #3 forces to inflexible distribution: Cut
4	CAUSEE Australia Number of full-time Employees Young Firms – Wave 4 (Year 4) Variable Name: W4_B16	Undetermined Lognormal or Weib	Undetermined Log or Weib	Undetermined Weib or LogN Lognormal	WeibvLogN: -1.48 (0.14)
5	CAUSEE Australia Number of full-time Employees Young and Nascent Firms – Wave 5 (Year 5) Variable Name: W5: Q24 – [NOTE**: Same variable than YF y NF – in row 10**	Undetermined Weib or Log or Cut	Undetermined Weib or Log or Cut	Cut – because inflexible criterion	CutvWeib: -0.73 (0.47) CutvLogN: -0.46 (0.65) WiebvLogN: 0.35 (0.72) No rejection of three: Cut, Weibull, LogN Rule #3 forces to inflexible distribution: Cut
6	CAUSEE Australia Number of full-time Employees Nascent Firms – Wave 1 (Year 1) Variable Name: W1: Q252#	Undetermined Lognormal (Between log or weib or Cut- To the limit in Cut 0.10)	Undetermined	Undetermined Log or Weib or even Cut	It seems lognormal but p value is high p=0.49 between Wieb y log CutvWeib: -2.19 (0.03) CutvLogN: -1.65 (0.1) WiebvLogN: -0.69 (0.49)
7	CAUSEE Australia Number of full-time Employees	Undetermined Lognormal	Undetermined Lognormal	Undetermined Lognormal	p=0.18 CutvWeib: -2.77 (0.17) CutvLogN: -2.07 (0.04)

	Nascent Firms – Wave 2 (Year 2) Variable Name: W2_C79	(and/or Weib)	(and/or Weib)	(and/or Weib)	WiebvLogN: -1.34 (0.18)
8	CAUSEE Australia Number of full-time Employees Nascent Firms – Wave 3 (Year 3) Variable Name: W3_C79	Undetermined Lognormal (and/or Weib)	Undetermined Lognormal (and/or Weib)	Undetermined Lognormal (and/or Weib)	P=0.24 CutvWeib: -2.05 (0.04) CutvLogN: -1.58 (0.11) WiebvLogN: -1.17 (0.24)
9	CAUSEE Australia Number of full-time Employees Nascent Firms – Wave 4 (Year 4) Variable Name: W4_C79	Undetermined Lognormal (and/or Weib)	Undetermined Lognormal (and/or Weib)	Undetermined Lognormal (and/or Weib)	P=0.31 Weibull v. lognormal: -1.0179 P= 0.3087253 CutvWeib: -2.96 (0.004) CutvLogN: -1.79 (0.07) WiebvLogN: -1.02 (0.31)
10	CAUSEE Australia Number of full-time Employees Young and Nascent Firms – Wave 5 (Year 5) Variable Name: W5_Q24 [NOTA MISMA VARIABLE QUE YF	Undetermined Weib or Log or Cut	Undetermined Weib or Log or Cut	Cut – because inflexible distribution Weib or Log or Cut	CutvWeib: -0.73 (0.47) CutvLogN: -0.46 (0.65) WiebvLogN: 0.35 (0.72) No rejection of three: Cut, Weibull, LogN Rule #3 forces to inflexible distribution: Cut
11	CAUSEE Australia Sales in $ (Total) (Last 12 Months) Young Firms – Wave 1 (Year 1) Variable Name: W1 Q2027#	Undeterminded Weib or Log or Cut	Undeterminded Weib or Log or Cut	Cut – because inflexible Weib or Log or Cut	CutvWeib: 54.16 (0.81) CutvLogN: 5.52 (0) WiebvLogN: -19.13 (0) No rejection of three: Cut, Weibull, LogN Rule #3 forces to inflexible distribution: Cut
12	CAUSEE Australia Sales in $ (Total) (Last 12 Months)	Cut	Cut	Cut	

	Young Firms – Wave 2 (Year 2) Variable Name: W2_B18				
13	CAUSEE Australia Sales in $ (Total) (Last 12 Months) Young Firms – Wave 3 (Year 3) Variable Name: W3_B18	Undetermined Weib or Log or Cut	Undetermined Weib or Log or Cut	Cut – because inflexible Weib or Log or Cut	No rejection of three: Cut, Weibull, LogN Rule #3 forces to inflexible distribution: Cut
14	CAUSEE Australia Sales in $ (Total) (Last 12 Months) Young Firms – Wave 4 (Year 4) Variable Name: W4_B18	Cut	Cut	Cut	
15	CAUSEE Australia Sales in $ (Total) (Last 12 Months) Young Firms – Wave 5 (Year 5) Variable Name: W5_Q18 [&R32] [note: same as NF]	Cut	Cut	Cut	
16	CAUSEE Australia Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 1 (Year 1) Variable Name: W1 Q2030#	Lognormal	Lognormal	Lognormal	Cut v Log, p= 0.0003065919
17	CAUSEE Australia Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 2	Lognormal	Lognormal	Lognormal	Cut vs Log; p= 0.05205102

	(Year 2) Variable Name: W2_C85_consolidated				
18	CAUSEE Australia Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 3 (Year 3) Variable Name: W3_C85	Undetermined Cut or Log Cut (o log?) p muy alto	Undetermined Cut or Log	Cut – because inflexible Weib or Log or Cut	Cut v Log norm LR: 0.8808874 p= 0.3783787 p value high
19	CAUSEE Australia Sales in $ (Total) (Last 12 Months) Nascent Firms – Wave 4 (Year 4) Variable Name: W4_C85_consolidated	Cut	Cut	Cut	
20	CAUSEE Australia Sales in $ (Total) (Last 12 Months) Nascent and Young Firms – Wave 5 (Year 5) Variable Name: W5_Q18[& R32] Misma variable que YF - 15	Cut	Cut	Cut	
21	SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave 1 (Year 0) Variable Name: gw31nn00	Undermined Lognormal or Weib	Undermined Lognormal or Weib	Undermined Lognormal or Weib	CutvWeib: -2.47 (0.01) CutvLogN: -1.82 (0.07) WiebvLogN: -0.81 (0.41)
22	SWEDISH PSED Number of full-time Employees– SWE PSED 1 Wave 2 (6 months)	Undermined Lognormal or Weib	Undermined Lognormal or Weib	Undermined Lognormal or Weib	WiebvLogN -1.54 (0.12)

	Variable Name: gw31nn06				
23	SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave 3 (12 months) Variable Name: gw31nn12	Undetermined Weib or Log or Cut	Undetermined Weib or Log or Cut	Cut – because inflexible Weib or Log or Cut	CutvWeib: 4.96 (0) CutvLogN: -0.58 (0.56) WiebvLogN: -0.58 (0.56)
24	SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave 4 (18 months) Variable Name: gw31nn18	Undetermined Lognormal or cut	Undetermined Lognormal or cut	Cut – because inflexible Weib or Log or Cut	PLvCut: -0.44 (0.35) PLvLogN: -0.76 (0.45) CutvWeib: - 6.85 (0.15) CutvLogN: -0.66 (0.50) WiebvLogN: -6.27 (0)
25	SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave 5 (24 months) Variable Name: gw31nn24	Undetermined Weib or Log or Cut	Undetermined Weib or Log or Cut	Cut – because inflexible Weib or Log or Cut	P values too high CutvWeib: -0.41 (0.68) CutvLogN: -0.39 (0.70) WiebvLogN: -0.39 (0.70)
26	SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave N75 (75 months) Variable Name: gw31n	Undetermined PL or Log or Cut	Undetermined PL or Log PL because rule #2	- inflexible distributions: the pure power law – rule #3 [PL or Log]	P values too high PLvCut: -0.51 (0.31) CutvWeib: 4.19 (0) PLvLogN: -0.82 (0.41) CutvLogN: -0.74 (0.46) WiebvLogN: -4.17 (0)
27	SWEDISH PSED - Outcome Variables Sales Turnover (Thousands SEK) Last Year Variable Name: pt11nn18	Lognormal	Lognormal	Lognormal	
28	SWEDISH PSED Sales Turnover (Thousands SEK)	Lognormal	Lognormal	Lognormal	

	First 3 Months Variable Name: pt12nn18				
29	SWEDISH PSED Sales Turnover (Thousands SEK) First 6 Months Variable Name: pt13nn18	Lognormal	Lognormal	Lognormal	
30	SWEDISH PSED Sales Turnover (Thousands SEK) First 12 Months Variable Name: pt14nn18	Undetermined Weib or Log or Cut	Undetermined Weib or Log or Cut	Cut – because inflexible Weib or Log or Cut	CutvWeib: 0.81 (0.42) CutvLogN: 1.26 (0.21) WiebvLogN: 0.51 (0.61)
31	SWEDISH PSED Sales Turnover (Thousands SEK) Second year of operation (24 months) Variable Name: pt11nn24 (global dataset)	Lognormal	Lognormal	Lognormal	
32	SWEDISH PSED Sales Turnover (Thousands SEK) Sales Turnover in 1997 Variable Name: pt31nn24 (global dataset)	Undetermined Lognormal (or Weib)	Undetermined Lognormal (or Weib)	Undetermined Lognormal (or Weib)	Small sample: 14 p values to high PLvWeib: 0.02 (0.99) CutvWeib: 1.16 (0.25) PLvLogN: -0.93 (0.35) CutvLogN: 1.91 (0.06) WiebvLogN: -0.71 (0.47)
33	SWEDISH PSED Sales Turnover (Thousands SEK) Sales Turnover in 1998 Variable Name: pt21nn24 (global dataset)	Undetermined Lognormal or cut	Undetermined Lognormal or cut	Cut – because inflexible Log or Cut	P value too high CutvLogN : 0.18 (0.86)
34	SWEDISH PSED	Lognormal	Lognormal	Lognormal	

	Sales Turnover (Thousands SEK) Last Year Sales Turnover after 75 months. Variable Name: pt11n (N75 SPSS file)				
35	SWEDISH PSED 35. Sales Turnover (Thousands SEK) Second year of operation (24 months) file SPSS erc-n24 Variable Name: SWE_pt11nn24_erc-n24	Lognormal	Lognormal	Lognormal	
36	36. Sales Turnover (Thousands SEK) Sales Turnover in 1998 Variable Name: pt21nn24_erc-n24 – ver otro file SPSS erc-n24 SWE_pt21nn24_erc-n24	Undetermined Lognormal or cut	Undetermined Lognormal or cut	Undetermined Cut – because inflexible Lognormal or cut	P values too high CutvLogN : 0.24 (0.81)
37	SWEDISH PSED Number of full-time Employees – SWE PSED 1 Wave 5 (24 months) Variable Name: gw31nn24 – **Specific dataset SPSS erc-n24**	Undetermined Weib or Log or Cut	Undetermined Weib or Log or Cut	Undetermined Cut – because inflexible Weib or Log or Cut	P values too high CutvWeib: -0.66 (0.51) CutvLogN: -0.54 (0.59) WiebvLogN: -0.07 (0.94)
38	SWEDISH PSED Sales Turnover (Thousands SEK) Sales Turnover in 1997 Variable Name:	Undetermined Lognormal (or Weib)	Undetermined Lognormal (or Weib)	Undetermined Lognormal (or Weib)	Very small sample PLvWeib: -0.20 (0.84) CutvWeib: 1.20 (0.23) PLvLogN: -1.10 (0.27)

	pt31nn24_erc-n24 – **Specific dataset SPSS erc-n24**				CutvLogN: 2.03 (0.04) WiebvLogN: -0.67 (0.50)
39	39. PSED II USA Total Revenues BV2	Lognormal	Lognormal	Lognormal	
40	40. PSED II USA Total Revenues CV2	Undetermined Lognormal or cut	Undetermined Lognormal or cut	Undetermined Cut – because inflexible Log or Cut	CutvLogN 0.64 (0.52) P value too high CutvWeib: 20.2 (0) CutvLogN: 0.64 (0.52) WiebvLogN: -6.29 (0.55)
41	PSED II USA 41. PSED II USA Total Revenues DV2	Undetermined Lognormal or cut	Undetermined Lognormal or cut	Undetermined Cut – because inflexible Log or Cut	Cut vs Log 0.52 (0.61) P value too high
42	42. PSED II USA Total Revenues EV2	Undetermined Weib or Log or Cut	Undetermined Weib or Log or Cut	Undetermined Cut – because inflexible Weib or Log or Cut	CutvLogN 0.70 (0.49) P value too high CutvWeib: 29.08 (0.83) CutvLogN: 0.70 (0.49) WiebvLogN: -6.75 (0.57)
43	43. PSED II USA Total Revenues FV2	Undetermined Lognormal or cut	Undetermined Lognormal or cut	Undetermined Cut – because inflexible Log or Cut	CutvLogN 0.60 (0.55) P value too high
44	PSED II USA Number regular	Undetermined	Undetermined	Undetermined	P too high

	Employees BU2 Variable: BU2	Weib or Log or Cut or Exp	Weib or Log or Cut or Exp Rule #2: Exp	Weib or Log or Cut or Exp Rule #3: Exp	PLvWeib: -2.60 (0.009) CutvWeib: -0.004 (0.99) PLvLogN: -2.44 (0.01) CutvLogN: 0.49 (0.62) WiebvLogN: 0.59 (0.55) CutvExp: 0.34 (0.74)
45	45. PSED II USA Number regular Employees CU2	Undetermined Weib or Log or Cut	Undetermined Weib or Log or Cut	Undetermined Cut – because inflexible Weib or Log or Cut	P too high PLvWeib: CutvWeib: -1.55 (0.12) PLvLogN: CutvLogN: -1.12 (0.26) WiebvLogN: -0.61 (0.54) CutvExp:
46	46. PSED II USA Number regular Employees DU2	Undetermined Weib or Log or Cut	Undetermined Weib or Log or Cut	Undetermined Cut – because inflexible Weib or Log or Cut	P high CutvWeib: -2.00 (0.045) CutvLogN: -1.19 (0.23) WiebvLogN: -0.17 (0.86)
47	47. PSED II USA Number regular Employees EU2	Undetermined PL or Log or Cut	Undetermined PL or Log PL v Cut: rule #2 PL nested	Undetermined PL or Log PL because is inflexible	P value high PLvCut: -0.41 (0.36) PLvLogN: -0.64 (0.52) CutvWeib: CutvLogN: -0.55 (0.59) WiebvLogN: -4.65 (0)
48	48. PSED II USA Number regular Employees FU2	Undetermined PL or Log or Cut	Undetermined PL or Log PL v Cut: rule #2	Undetermined PL or Log PL because is	P value high PLvCut: -0.29 (0.45) PLvLogN: -0.43 (0.67) CutvWeib:

			PL nested	inflexible	CutvLogN: -0.27 (0.78) WiebvLogN: -5.17 (0)